Ecological genomics and speciation in malaria vector mosquitoes

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

Submitted by

Chris S. Clarkson

December 2015

0.1 Acknowledgements

First, and foremost, I would like to thank my supervisors, Martin Donnelly and David Weetman. Without their guidance, encouragement, idioms and occasional inappropriate use of fonts, my PhD would have been a much more difficult and a far less enjoyable experience; Gordian knots would certainly not have been cut and I would not be on the pig's back now. Tiago Antão's presence in the group, with his expert knowledge on all things computer and population genomics, also had a huge effect on this thesis. I would like to thank him for his patience in dealing with my many questions about bioinformatics and letting me bounce my ideas for analyses off him.

Thanks must also go to those whose skills in the field, laboratory and behind computers this project has benefitted from greatly. Ghanaian fieldwork would not have been possible without the help Alexander Egyir-Yawson and the endless hard work, above and beyond the call of duty, of John Essandoh. My mosquito colonies would certainly have not survived without babysitting services of Adriana Adolfi and Amy Lynd or the insectary skills of Jonathan Thornton and life in in the lab was made much easier via Keith Steen and Emily Rippon's help and advice. I was lucky enough to have access to high quality genomic data throughout my PhD, for this I have the sequencing teams at the Wellcome Trust Sanger Institute and the Broad Institute to thank.

Outside of LSTM there are some people who deserve a special mention. All the London folk, who have supported me through the stressful times, fixed broken code and understood my absences, and Amy, who gets a second mention for being my favourite landlady and a great friend. I must also thank the people that have inspired and pushed me to get this far, my wonderful parents for putting up with having a perpetual student for a son, Nat, Judy and Ruth, simply for putting up with me. Finally, for the grandparents I lost during my final year, who gave me 32 years of love and support, I dedicate this thesis to them, Mary and Peter Field.

0.2 Abstract

Malaria vector control campaigns generate rapid changes to ecology with concomitant selective pressures on their targets. Malaria mosquitoes, therefore, present an excellent system for studying ecological adaptation alongside the medical importance of understanding the evolution of insecticide resistance.

Adaptive introgression allows rapid adaptation through gene flow of fitness conferring loci from another species and, though commonly reported in plants, it is thought to be rare in animals. Using a whole genome sequencing approach, the impact of an insecticide resistance locus introgression between malaria vectors, *A. gambiae* and *A. coluzzii*, was investigated. The resistance locus lies within a genomic 'island', a large region of genome, highly divergent between the two species. Our results revealed that the inter-specific introgression transferred, not only the *kdr* resistance locus, but ~1.5% of the surrounding genome from *A. gambiae* into *A. coluzzii*, homogenising the entire island. These findings bring into debate hypotheses of the genomic islands' role in the speciation of these mosquitoes, as no increase in hybridisation had been recorded despite the large introgressed region sweeping through the *A. coluzzii* population. From a control perspective, these results also demonstrate how quickly a species can react to anthropogenic pressure, to escape chemical control.

In Ghana, as across much of *A. gambiae* and *A. coluzzii*'s sympatric range, low levels of hybridisation are recorded, a fact that was in part used to validate the recent elevation to specific status of the pair. However, in the far-west of their range, much higher levels of hybridisation are reported. A microsatellite study, sampling a transect across a region known for high gene-flow in Guinea Bissau, was carried out to investigate this phenomenon, discovering a hybridisation hot-spot in the coastal region. We augmented this microsatellite study with WGS data, sequencing individuals from coastal and inland sites. Results revealed that in the coastal region alongside molecularly genotyped hybrids, mosquitoes which had been genotyped as pure species appeared to be admixed. Ancestry informative markers demonstrated asymmetric introgression from *A. coluzzii* into the coastal *A. gambiae* and, with *A. coluzzii* numbers falling in the region, species collapse or perhaps hybrid speciation may be occurring. In these high gene flow regions the mosquitoes do not conform to distinct taxonomic units and we show that whole genome characterisation is necessary to understand the evolutionary dynamics.

Understanding insecticide resistance is a major motivation in researching the genomics of malaria vectors, as resistance threatens the success of control campaigns. The voltage gated sodium channel (VGSC), is of particular interest, carrying several loci known to confer insecticide resistance in *A. gambiae* and *A. coluzzii*. We utilised 765 WGS individuals from the *Anopheles gambiae* 1000 Genomes Project, to investigate gene flow and evolution in this important gene. The data was phased then displayed in a network approach, to enable relationships between the VGSC haplotypes to be discerned. Striking was both the evidence for an abundance of long range gene flow, with resistant haplotypes shared across vast geographical distances, and for the high numbers and placement of non-synonymous mutations on haplotypes carrying resistance mutations. A hitherto unknown complexity of both the spread of resistance and protein altering mutations was found, suggesting current molecular assays may need revising.

Insecticide resistance is perhaps less well understood in another major malaria vector, *A. arabiensis*. Recent advances in genomic resources for the species enabled us to carry out a genome wide association study (GWAS) to identify insecticide resistance candidates. To produce confidence in results, a pool-seq approach was taken, sequencing and comparing the allele frequencies of over 1000 resistance phenotyped individuals. Several candidate regions were found, the strongest of which we investigated further. The ~225kb region on the 2R chromosome arm straddled a cluster of cytochrome P450 genes, known to be involved in detoxification, including *CYP6P2*, a gene previous linked to pyrethroid resistance in this species. These results revealed this technique to be both viable and useful for identifying phenotype/genotype candidates in wild populations of mosquitoes.

The work herein demonstrates that the genomic study of *Anopheles* malaria vectors is not just powerful in medical context, but that these animals provide excellent natural systems for evolutionary study. *A. gambiae* in Ghana and in Guinea Bissau provide natural experiments contrasting how species integrity reacts in low and high gene flow situations, while investigations into the VGSC and insecticide resistance in *A. arabiensis* shine light on how organisms adapt to rapidly changing ecologies.

0.3 Table of contents

Chapter 2 Adaptive introgression eliminates a major genomic island of divergence but not reproductive isolation between *Anopheles gambiae* sibling species

Chapter 3 Species collapse in the "Wild West"? Genomic replacement by asymmetric introgression in an *Anopheles* hybrid zone

Chapter 4 Evolution of the *Anopheles gambiae* voltage gated sodium channel gene: a haplotype network approach

0.4 Figures

0.5 Tables

0.6 Abbreviations

Ag1000G – *Anopheles gambiae* Thousand Genomes Consortium

AIM – ancestry informative markers

CA - carbamates

cDNA – copy deoxyribonucleic acid

CI – confidence interval

CNV – copy number variant

CYP450 – cytochrome P450

DDT - dichlorodiphenyltrichloroethane

DNA – deoxyribonucleic acid

DP – depth

FS – Fisher strand

ftp – file transfer protocol

GQ – genotype quality

GWAS – genome wide association study/studies

Hrun – homopolymer run

IRS – indoor residual spraying

ITN – insecticide treated bednet

kdr – knock down resistance

LD – linkage disequilibrium

MAF – minor allele frequency

MQ – mapping quality

MYA – million years ago

OP - organophosphates

PC – principal component

PCA – principal component analysis

PCR – polymerase chain reaction

QD – quality by depth

QTL – quantitative trait locus/loci

RAD – restriction association digest

rDNA – ribosomal deoxyribonucleic acid

RI – reproductive isolation

RNA – ribonucleic acid

RSI – repetitive strain syndrome

SD – standard deviation

SGF – speciation with gene flow

SINE – short interspersed nuclear element

SNP – single nucleotide polymorphism

VCF – variant call format

VGSC – voltage gated sodium channel

WGS – whole genome sequencing

Chapter 1

Genomics of adaption and speciation: Review of the literature

1.1 Introduction

Since the publication of Darwin's most famous tome, The Origin of Species (Darwin, 1859), the biological debate about speciation has been a lively one. Ernst Mayr's research in the 1940s led to geographic isolation being seen as de-rigueur for speciation to occur (Mayr, 1942). The argument being, only in allopatry could species evolve reproductive isolation (RI) mechanisms such as those advocated by Dobzhansky (1934; 1937) and Muller (1942). Due to this prevailing orthodoxy and later, Kimura's popular neutral theory suggesting directional selection on adaptive mutations was rare (Kimura, 1984), sympatric speciation (or adaptation with gene flow) was seen as unlikely and somewhat contentious. The advent of the genomic era has re-opened this debate and provided evidence that sympatric speciation both can and does occur. One of the major models for studying speciation with gene flow are the malaria vector sibling species, *Anopheles gambiae* and *Anopheles coluzzii* (Turner, Hahn and Nuzhdin, 2005; Lawniczak *et al*., 2010; Neafsey *et al*., 2010; Fontaine *et al*., 2015).

1.2 Speciation with gene flow

As the geographic relationship between populations often affects gene flow between them, these relationships are used to define modes of divergence and speciation (Mayr 1963; Bolnick and Fitzpatrick, 2007; Nosil, 2008). At one extreme, allopatric speciation sees diverging populations physically separated with no gene flow, a mode once thought to be dominant (Mayr, 1963). At the other extreme, sympatric speciation, "with-out geographic barriers" (Mayr, 1963) between diverging populations where "mating is random with respect to the birthplace of the mating partners" (Gavrilets, 2003) allows free gene flow between speciating entities. This latter model is/was thought to be unlikely due to the homogenisation effects of gene flow on divergence (Coyne and Orr, 2004). Intermediate scenarios also exist, such a parapatric speciation, with partial extrinsic barriers to gene flow (Butlin, Galindo and Grahame, 2008).

Despite theoretical studies revealing how and when divergence may occur in sympatry (*e.g.* Smith 1966; Barton and Charlesworth, 1984; Rice and Hostert, 1993; Dieckmann and Doebeli, 1999) and empirical evidence (see below), criteria developed for identifying truly sympatric speciation were not met and it was still thought as being extreme and uncommon (Coyne and Orr, 2004; Bolnick and Fitzpatrick, 2007). However, with the cost of whole genome sequencing (WGS) dropping drastically (Baker, 2010) and cheaper genomic techniques such as restriction associated digest (RAD) being developed (Baird *et al*., 2008), recent years have seen many high resolution genome-scale empirical studies finding evidence of speciation with gene flow (SGF) or adaptation with gene flow in many study systems. In concert with this wealth of new data, new theoretical frameworks, such as Wu's mosaic genome, were developed and have been exploited with genome scale data (Wu, 2001). Along with these new resources was the acknowledgement that satisfying the demands of Mayr's sympatric speciation criteria (1963), are likely to be impossible without knowledge of a population/species demographic history and these data are often difficult to estimate (Coyne and Orr, 2004). Whilst the direction of research may not have altered, our terminology has, with the broader 'speciation-with-gene-flow' (SGF) nomenclature supplanting the "relatively uncommon" sympatric speciation (Bolnick and Fitzpatrick, 2007; Nosil, 2008).

1.3 Pre-genomics empirical evidence for SGF

Despite being a controversial topic, empirical evidence for SGF existed pre-whole genome science. One established evolutionary pathway for the establishment of reproductive isolation (RI) with gene flow is via host races. Here populations of organisms with strong trophic relationships such as phytophages or parasites become adapted to feeding on different species of host (Diehl and Bush, 1989). Though these host races may be geographically close and able to exchange genes, the specialisation to their host results in reductions in hybrid fitness and the evolution of RI mechanisms. One of the most widely studied of these systems is *Rhagoletis pomonella*, an apple feeding race evolving from the ancestral hawthorn feeding race in the last ~150 years (Feder, 1998; Powell *et al*., 2013). The adaption of the two races to the differing phenology of hosts was found to be driving isolation into nascent species (Feder *et al*., 1997; Powell *et al*., 2013). Another well studied system, the pea aphid host races of alfalfa and red clover (*Acyrthosiphon pisum*), were found to exhibit a similar evolutionary path, though here the habitat choice of the winged

colonisers was defining the races, with pre- and post-zygotic RI through selection against migrants and hybrids (Via, Bouck and Skillman, 2000).

Ecological diversification leading to sympatric divergence had not just been noted in host races. In African crater lakes, with no connection to river systems, numerous species of cichlid fishes have been found (Schliewen, Tautz and Pääbo, 1994). Molecular analysis of mitochondrial DNA revealed monophyly of each lake's species and suggested a single introduction of fish into each lake, followed by divergence into multiple species through habitat specialisation (Schliewen, Tautz and Pääbo, 1994). Rather than host plant niches the fish appear to have diversified to fill roles at different trophic levels. Extremes of sympatry however, like the closed cichlid lake system, are not necessary for the generation of non-allopatric speciation evidence. Parapatric divergence, as discovered between the morphs of *Littorina saxatilis*, has also revealed that it is possible for overlapping and thus gene sharing populations to diverge. When 306 AFLP markers were compared across low and high shore *L. saxatilis* a small number were found to differ more than expected due to selection (Wilding *et al*., 2001). These results fitted a model of SGF where loci under disruptive selection can diverge but other loci cannot due to continued gene flow (Rice and Hostert, 1993; Wilding *et al*., 2001).

1.4 Theoretical advances – the mosaic genome

An important theoretical pre-cursor that enabled an interpretation of the empirical genomic SGF results to come was Wu's genic view of speciation (2001). Mayr's biological species concept (*sensu stricto*) was based on 'good species' having complete RI (Mayr, 1963), therefore isolation functioned on a whole organism or whole genome level. However, Wu noted that evidence suggests selection acts at a finer resolution, on a genic level. This allowed for the hypothesis of a mechanism where divergence and speciation could progress in the presence of gene flow. One or more loci under divergent selection, *e.g.* disruptive selection, between populations otherwise freely sharing genes would be protected from gene flow by fitness effects. As introgression at these selected loci would be maladaptive, effective gene flow is reduced. Through linkage to selected loci additional loci are captured. These may not initially be under strong selection but low recombination allows them to diverge, generating co-adapted gene complexes and further loci contributing to divergence, expanding the effect of reduced recombination. Genome wide, gene flow continues but in a mosaic fashion, and decreases as the divergent regions grow. Eventually RI mechanisms

may curtail all gene flow resulting in two 'good' species (Figure 1.1) (Wu, 2001; Wu and Ting, 2003).

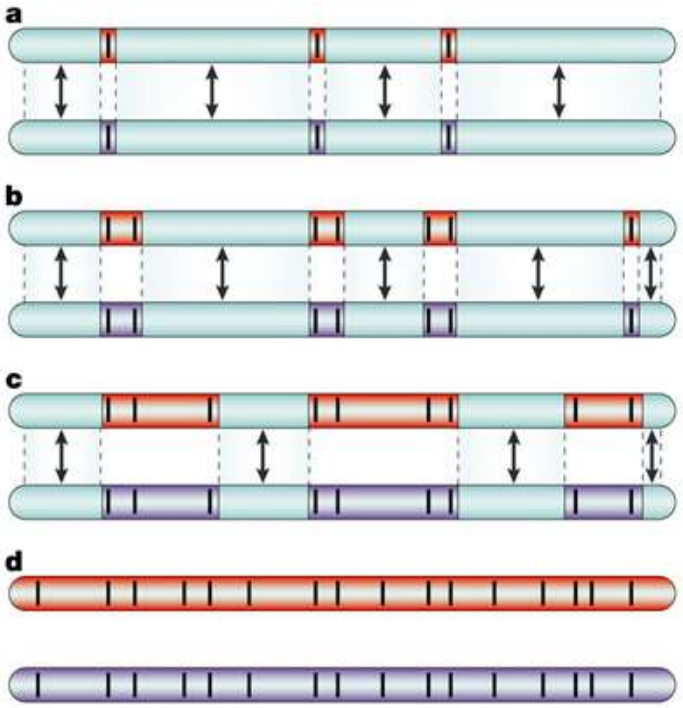


Figure 1.1. The genic view of species differentiation. The two bars represent genomes of two diverging populations. (**a**) When they start to differentiate, only a few loci (black lines) are differentially adapted and genes at these loci are not exchanged between populations. Gene flow continues (black arrows) in the rest of the genome. (**b**, **c**) The regions of differential adaptation expand, the amount of gene flow between the two genomes is gradually reduced owing to linkage with such regions (indicated in red/purple) until, (**d**) the two populations are completely reproductively isolated and are therefore considered to be separate species. Adapted from Wu, 2001; Wu and Ting, 2004.

1.5 Genomics of SGF – empirical evidence

If one empirical paper could be said to have spurred the recent wave of investigations into the genomics of speciation, it is the paper on incipient speciation in the major malaria vector mosquito, *Anopheles gambiae* (Turner, Hahn and Nuzhdin, 2005). The paper investigated *A. gambiae*'s then two molecular forms (now species – Coetzee *et al.*, 2013), which are commonly sympatric and thought to be diverging based on differences in larval habitat

ecology (Gentile *et al*., 2001; Lehmann and Diabeté, 2008). Assortative mating is observed in natural settings (Tripet *et al*., 2001; Tripet *et al*., 2003), and with no apparent fitness costs found in hybrids in the lab (Diabeté, 2007), this divergence was thought to be recent and caught in process (Turner, Hahn and Nuzhdin, 2005). Using an approach based on microarray technology, the authors hybridised genomic DNA from the two forms (rather than cDNA used in transcriptomic studies) to a microarray to quantify divergence across the *A. gambiae* genome. What was striking about the results was how closely they fitted Wu's mosaic genome model for speciation with gene flow (2001). Against a background of low genome-wide divergence between the forms, as may be expected in early stages of SGF, three divergent regions ("islands") were discovered and were shown to contain fixed differences, which it was argued were unlikely caused by neutral processes. The most highly divergent of these genomic islands were adjacent to centromeres of the 2L and X chromosome arms and the authors concede divergence in these regions is possible without directional selection, owing to low background recombination rates (Turner, Hahn and Nuzhdin *et al*., 2005; Carneiro, Ferrand and Nachman, 2008). However, it was suggested that the loci driving the ecological divergence seen between the forms could be contained within these three divergent islands and that this technique could be used to locate the 'holy grail' of speciation research, speciation genes (Turner, Hahn and Nuzhdin, 2005).

The technique used by Turner *et al*. (2005) has since been developed and augmented by the use of WGS data as it became more readily available. This increase in resolution with millions of markers allowed a variety of divergence and selection detection statistics to be applied in scans across the genomes of diverging populations to detect candidate regions for speciation. One of the first WGS studies was conducted on flycatchers (Ellegren *et al*., 2012). The recently diverged (<2 MYA) collared flycatcher (*Ficedula albicollis*) and the pied flycatcher (*Ficedula hypoleuca*) show pre- and post-zygotic isolation, however, they hybridise where populations are sympatric. Using a suite of statistics, around 50 highly divergent islands were detected, with concordant changes across multiple related and orthogonal statistics; increases in the density of fixed differences (d_f - skewed allele frequency spectrum), reductions in π (nucleotide diversity), expected heterozygosity (Tajima's D) and increases in LD (Figure 1.2). It should be noted that although calculated, there is no divergent island signal in D_{xy}, another independent statistic, which has been suggested to be a key to identifying 'true' islands, something that will be discussed later (see **Islands in hot water**). As with the mosquitoes, many genomic islands between the flycatchers appeared to lie in regions where centromeres and telomeres were predicted and

again the authors suggested that these structural features may be involved in species divergence (Ellegren *et al*., 2012).

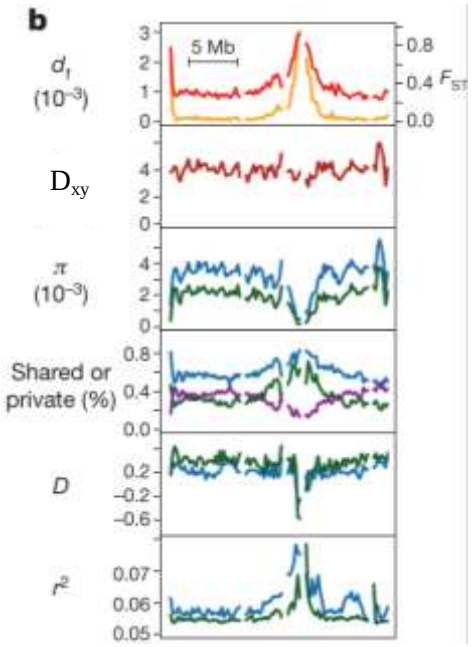


Figure 1.2. Distribution of population genomic parameters along flycatcher chromosome 4A. Chromosome 4A was chosen as an example presenting a divergent genomic island. The density of fixed differences per base pair across the chromosome (d_f) is shown in yellow, F_{ST} in red and D_{xy} shows the total between species divergence (brown). For nucleotide diversity (π), proportion of private polymorphisms, Tajima's D and linkage disequilibrium (r^2), estimates across the chromosome for collared flycatcher are shown in blue, pied flycatcher in green. The proportion of shared polymorphism is shown in purple. Note the absence of signal in D_{xy} compared to all other metrics. Adapted from Ellegren *et al*., 2012.

The use of genomic techniques like these to identify genomic islands of divergence has, in just ten years since Turner *et al*. (2005) first identified them, revolutionised the study of speciation. Application to many systems, where incipient SGF is thought to be occurring or where divergence occurred recently, has identified regions of the genome (genomic islands) that appear to be under divergent selection (*e.g.* Nadeau *et* al., 2012; Karlsen *et al*., 2013, Ruegg *et al*., 2014). What are less common, however, are studies which actually demonstrate that loci within these islands are driving the phenotypic differences between the populations or species in question, that the islands are not just artefacts of genome structure. Research

into the three spined stickleback (*Gasterosteus aculeatus*) freshwater and marine populations showed that some of the divergent islands found were located in QTLs previously identified as being involved in phenotypic differences (Hohenlohe *et al*., 2010), but it was research into crow species that perhaps best illustrates the power of these techniques.

The study set out to characterise the genomic differentiation between carrion and hooded crows, which form stable hybrid zones across Europe where populations overlap (Poelstra *et al*., 2014). Using WGS data and similar techniques to the other studies discussed, the authors performed genome scans and found one genomic island region, representing less than 1% of the genome against a background of low divergence, suggesting generally high gene flow between the species. Within the island, reduced nucleotide diversity was found alongside increased LD and 81 of the 82 observed fixed differences between species, providing strong evidence that this region was involved in species differentiation. Interestingly two F_{ST} peaks connected by a saddle of high F_{ST} suggested that this region may be an inversion, another structural feature of the genome, like centromeres and telomeres, known to affect recombination and divergence (Kirkpatrick and Barton, 2005; Carneiro, Ferrand and Nachman, 2008). What makes this study so exciting is that the authors also performed RNA sequencing analysis which demonstrated that genes within the island were also being expressed differently between the species, genes not only involved in plumage colouration pathways (explaining phenotype differences) but also genes involved in visual pathways (possibly accounting for the assortative mating found between the species) (Poelstra *et al*., 2014). The study functionally supported the mosaic genome and islands of speciation theories (Wu, 2001; Turner *et al*., 2005) but suggested that structural features of the genome, like inversions, may be important for supporting creation of co-adapted gene complexes ("supergenes") in the speciation process (Joron *et al*., 2006).

1.6 Relevance of studying speciation with gene flow

Historically, the study of speciation has been concentrated on RI mechanisms, for example Muller and Dobzhansky's work on hybrid sterility and viability (Dobzhansky, 1934; Muller, 1942). Though retrospective analysis of speciation has provided data on isolating mechanisms, the populations under scrutiny may be thousands or millions of years diverged, leading to difficulties in discerning the initial conditions and drivers of speciation. Via describes this traditional approach as the "spyglass"; attempting to look back in time at speciation but with confounding effects of population histories reducing the accuracy of our

inferences with increasing time (Figure 1.3A) (Via, 2009). The recent trend for the study of populations diverging but still undergoing gene flow however (what Via calls the "magnifying glass"), allows for isolating mechanisms to be identified with few confounding factors (Figure 1.3B) and if populations are caught early enough in their divergence, ancestral populations may still be available for sampling. The approach also allows, funding permitting, for diverging populations to be followed through time to see how divergence/RI/speciation progresses (Via, 2009). How the populations' ecology interacts with isolation can also be measured rather than guessed. By studying populations that are not fully isolated we can ask how speciation occurs, not just what do species look like.

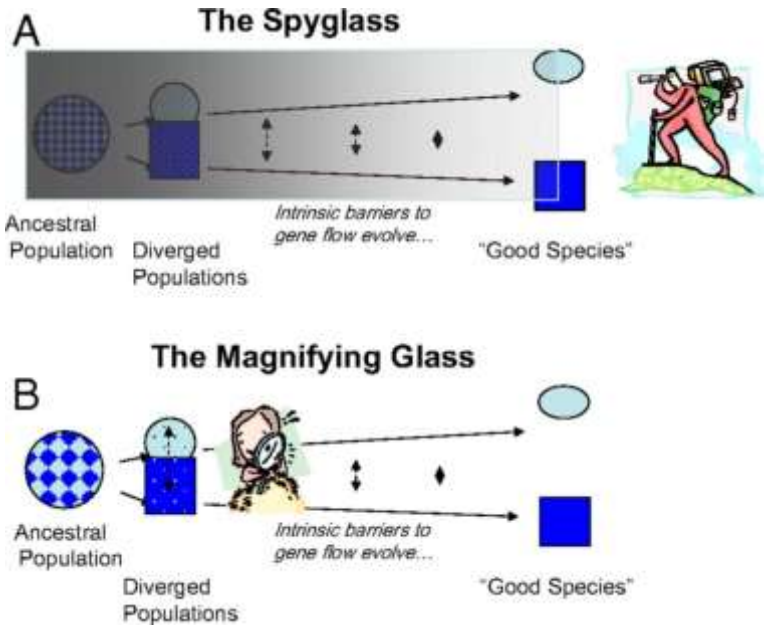


Figure 1.3. Two ways of studying speciation. Vertical arrows represent gene flow, reducing as divergence increases and intrinsic barriers to gene flow evolve (horizontal arrows) until "Good Species" are fully reproductively isolated. (**A**) The Spyglass; starting with good species and trying to look backwards in time to discern drivers of divergence. The increase in shading going back in time represents the increase in difficulty in predicting how isolation evolved due to confounding demographic factors. (**B**) The Magnifying Glass; by finding populations early in the divergence and studying divergence drivers or how isolating mechanisms evolve, confounding factors are reduced and ancestral populations may still be extant and available for comparison. Adapted from Via, 2009.

1.7 Theoretical advances – the species continuum

With an abundance of empirical SGF data being produced, theoretical models had substrates to test their results against and advances began to be made towards better understanding the generation and progression of the divergence landscape during SGF. Models were developed to show how genomic islands appear, increase in size and ultimately how RI can evolve in the face of homogenising gene flow. These new ideas and corresponding terminology are explored below.

1.7.1 -island metaphor

The metaphor of islands given by Turner *et al.* in their seminal *Anopheles* paper (2005) has since been expanded to help explain the sequence of genomic events that are thought to take place during SGF. The sea level is the threshold of neutral expectation, outliers beyond this forming the genomic islands, potentially composed of loci under divergent selection and neutral markers in linkage with them, selection strength and recombination therefore, both affecting island size (Michel *et al*., 2010; Nosil and Feder, 2012). The island/sea genomic mosaic allows this divergence to occur in the face of homogenising gene flow (Wu, 2001; Wu and Ting, 2004).

1.7.2 -divergence hitchhiking

Genomic islands arise and increase in size by divergence hitchhiking. When divergent selection acts upon a locus it reduces gene flow of the locus between populations, this has the effect of reducing the effective recombination rate around the locus. Reduced recombination allows the region around the selected locus to diverge between population both by purely neutral processes and by helping establish other divergent loci (potentially co-adapted with the initial loci), causing the genomic island to increase in size (Via, 2012; Nosil and Feder, 2012) (Figure 1.4 – the island view of divergence).

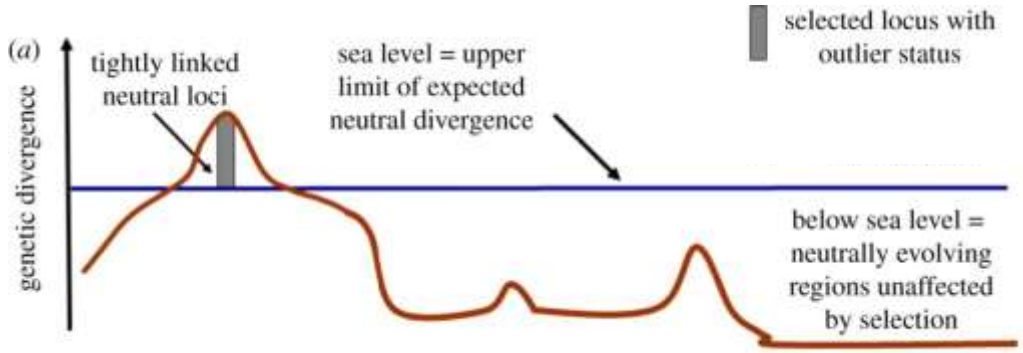


Figure 1.4. Schematic of the island view of divergence – divergence hitchhiking. Neutral loci in close linkage to the loci under directional selection diverge through reduced recombination and increased effects of genetic drift Adapted from Michel *et al.*, 2010.

1.7.3 -genomic hitchhiking

Keeping with the island metaphor, as speciation progresses, divergence will be found across more of the genome, the genomic islands grow into archipelagos then continents (Figure 1.5). With continued divergence, genomic islands increase in size and number, this extends the localised reduction in gene flow and recombination to the entire genome allowing it to rise above 'sea level' divergence. This is known as genomic hitchhiking (Feder *et al.*, 2012).

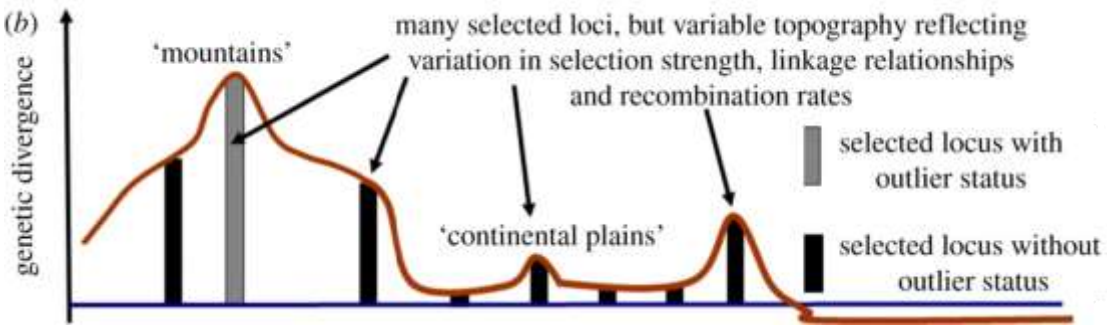


Figure 1.5. Schematic of the continent view of divergence – genomic hitchhiking. As reduced recombination allows more weakly selected loci to diverge the islands grow in size, becoming "continents". Adapted from Michel *et al.*, 2010.

1.7.4 -the species continuum

Rather than alternative views, the island and continent views of speciation are just points on the SGF continuum (Nosil and Feder, 2012); genomic continents forming as islands rise from the neutral sea level and grow in size. The SGF theory regarding how divergence is expected to progress allows predictions to be made regarding how far along the speciation continuum populations under analysis are. Four stages of the species continuum have been proposed to capture the changes in topography of divergence as speciation progresses. In phase 1, direct selection on a locus causes differentiation at this locus and neutral, closely physically linked sites near to it, then through divergence hitchhiking the size of this genomic island increases (phase 2) (Figure 1.6). In phase 3, as the increased effects of selection decrease recombination across the whole genome and unlinked neutral loci begin to diverge until, in phase 4, the whole genome is isolated regardless of LD with the locus initially under selection (Figure 1.6) (Feder *et al*., 2012). It should be noted that progression through the species continuum will not proceed at a uniform rate from instance to instance and also that it is not unidirectional with 'good' reproductively isolated species being the inevitable end point (Feder *et al*., 2012). For example, if the selective environment changes, population divergence could begin to reverse, resulting in species collapse or hybrid swarm events rather than speciation (Pritchard and Edmands, 2013). Alternatively, gene flow could persist indefinitely creating a stable mosaic genome.

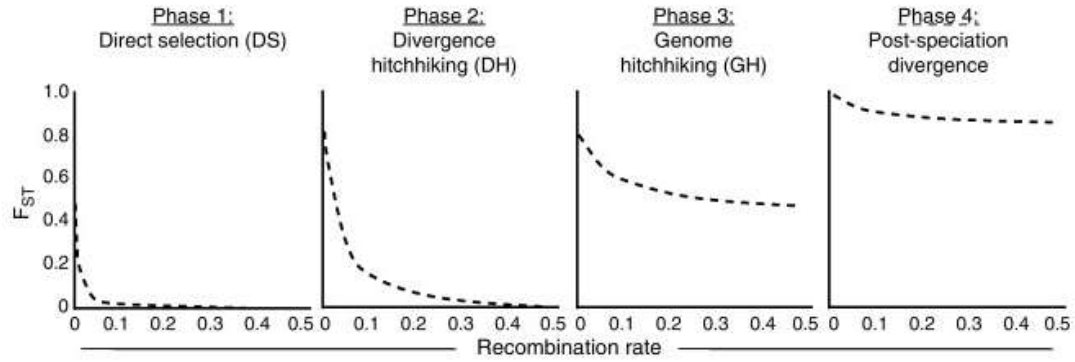


Figure 1.6. Four potential stages of speciation with gene flow. The expected relationship of divergence at neutral sites (measured with F_{ST}) with respect to recombination rate, from completely linked to the divergently selected locus (r = 0cM) to completely unlinked (r = 0.5cM), as speciation progresses. Phase 1: divergence is only found at, or very closely linked to, the locus under divergent selection. Phase 2: divergence hitchhiking allows divergence close to the selected locus to increase and for less linked loci nearby to diverge. Phase 3: recombination across genome is reduced allowing even unlinked sites to diverge - genomic hitchhiking. Phase 4: Reproductively isolated species are free to diverge by both neutral and selective forces. Adapted from Feder, Egan and Nosil, 2012.

1.8 Islands in hot water?

In 2005, Turner *et al*. noted that using genomic islands of divergence to identify candidate regions as drivers of SGF was complicated by the fact these islands were often in or near genomic structural features that can reduce recombination (Turner *et al*., 2005 and commentary by Butlin, 2005), such as centromeres (Carneiro, Ferrand and Nachman, 2008) and inversions (Kirkpatrick and Barton, 2006). Low within-population diversity in these regions can result in highly divergent loci between populations in the absence of directional selection (Charlesworth, 1998). These signals are difficult to untangle, particularly as low recombination regions may actually aid the formation of co-adaptive speciation gene complexes. Low recombination regions may also allow retention of signals, for example those that may have been important in early stages of divergence of selection but for which selection coefficients have since reduced, for long enough that they can be detected before being obscured by recombination (Butlin, 2005).

Noor and Bennett, in their provocatively titled paper, ask if they really are "islands of speciation or mirages in the desert?" and note three main issues (2009). Firstly genomic islands are often found within inversions (*e.g.* Michel *et al*., 2010), however genetic divergence found in these regions could be due to the inversion divergence being older than the population divergence (neutral forces causing nucleotide divergence) as found in *Anopheles* species (White *et* al., 2009), or contrastingly, because inversions can spread, post divergence, through directional selection leaving divergence which may have nothing to do with SGF (Kirkpatrick and Barton, 2006). Secondly, low recombination regions may appear more diverged between populations because of recurrent hitchhiking or background selection (Charlesworth, 1998) and QTLs are easier to map to low recombination regions due to stronger marker-QTL linkage (Noor and Bennett, 2009). Thirdly, even the initial premise that populations are actually exchanging genes may be confounded by incomplete lineage sorting leaving shared alleles between isolated populations (Hey, 2006; Noor and Bennett, 2009), something later acknowledged may be the case in the *A. gambiae/coluzzii* system (White *et al*., 2010; Turner and Hahn, 2010). However, Noor and Bennett (2009) specified that only some of the then-extant empirical results appeared to be at risk from these problems, whilst cautioning that the statistical techniques for linking genomic islands with speciation must be applied with caution.

One point that is repeated throughout these cautionary commentaries is that the use of relative measures of divergence, for example F_{ST}, may not be the most appropriate statistics (Charlesworth, 1998; Noor and Bennett, 2009). Pre-genomics, it had already been noted that F_{ST} was strongly affected by intra-population genetic diversity (Charlesworth, 1998). Regions of the genome with low recombination and therefore often low within-population diversity may elevate F_{ST} in the absence of divergent selection (Charlesworth, 1998). Relative measures of diversity are affected as they measure the proportion of between population differentiation relative to the overall diversity. Charlesworth recommended the use of absolute measures of divergence, those which are independent of population diversity, however this appears to have gone unheeded and many studies of SGF still used only relative measures (*e.g.* Stump *et al*., 2005; Turner, Nuzhdin and Hahn, 2005; Slotman *et al*,. 2006; Carneiro, Ferrand and Nachman, 2008). Noor and Bennett (2009) reiterated Charlesworth's point (1998), taking data from Stump *et al*. (2005) to show that results were very different when the absolute measure of divergence D_{xy} was used (Nei, 1987).

A more thorough analysis of relative versus absolute measures using previously published data was conducted by Cruickshank and Hahn (2014), who took five species pairs from previously published studies on divergence with gene flow. Separating the data into genomic island and non-genomic island regions using F_{ST}, D_{xy} was then calculated for these regions. Results revealed that absolute divergence in these island regions was not elevated; in fact it was often lower in islands than non-islands. The authors also found island regions to be lower in nucleotide diversity than non-islands (as expected with many islands being found in low recombination regions) and suggest it was in fact this causing the elevated F_{ST} rather than directional selection followed by reduced gene flow between populations (Charlesworth, 1998). To explain the presence of reduced diversity within islands a 'divergence after speciation' model is used where the species pairs are no longer exchanging gene flow and selection (in isolation) causes divergence; signals of this are most likely to be retained in regions of low recombination (Figure 1.7) (Cruickshank and Hahn, 2014). However, D_{xy} suffers from high variance with low sample sizes (which was the case in the studies analysed) alongside high stochastic variance between variants (Wakeley, 1996). This creates problems for interpretation of 'negative' results and perhaps an improved absolute metric is called for. Indeed, it is interesting to note that there is empirical SGF evidence with a genomic island identified using relative divergence metrics where genes within the island region were suggested to be driving divergence between the species yet no signal was found in region using D_{xy} (Poelstra *et al*., 2014).

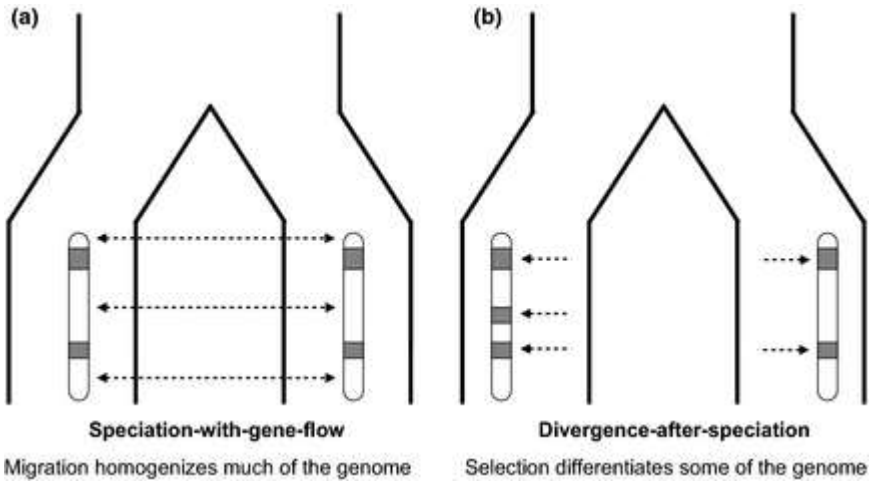


Figure 1.7. Speciation with gene flow model vs. Divergence after speciation model. (**a**) Arrows represent gene flow and divergent regions resistant to gene flow are shown in grey. (**b**) Arrows show where selection has driven divergence (shown by grey band) and white shows where daughter species are similar due to shared ancestry and incomplete lineage sorting. Adapted from Cruickshank and Hahn, 2014.

1.9 Mosquitoes as a tractable system for studying speciation with gene flow

The system that kick-started the study of the genomics of SGF (Turner, Hahn and Nuzhdin, 2005), *A. gambiae*, is particularly tractable for the study of these evolutionary phenomenon. Formerly the molecular forms of *A. gambiae*, M and S, the now sister species *A. coluzzii* (M) and *A. gambiae* (S) (Coetzee *et al.*, 2013) are thought to have only recently diverged within ~0.54 million years (Fontaine *et al.*, 2014) and have a wide range with large areas of sympatry across Sub-Saharan Africa (Figure 1.8). These species display a range of reproductive isolation from each other. Hybridisation rates of less than1% are usually found in sympatry (della Torre, Tu and Petrarca, 2005, Simard *et al.*, 2009; Tripet *et al.*, 2001), however in The Gambia, 7% of mosquitoes caught were hybrids while in Guinea Bissau more than 20% of wild caught females were putative hybrids (Caputo *et al.*, 2008; Oliveira *et al.*, 2008). Ranges of hybridisation and resultant gene flow (Weetman *et al.*, 2012), allow different stages of divergence to be investigated over tractable temporal scales. Despite their elevation to specific status (Coetzee *et al.*, 2013), no post mating isolation is found in the lab (Diabaté *et al.*, 2007) and the sister species show clear signs of contemporary gene flow even in regions with low levels of hybridisation; an insecticide resistance locus being shown

to have introgressed between the species in recent years (Weill *et al*., 2000; Weetman *et al*., 2012).

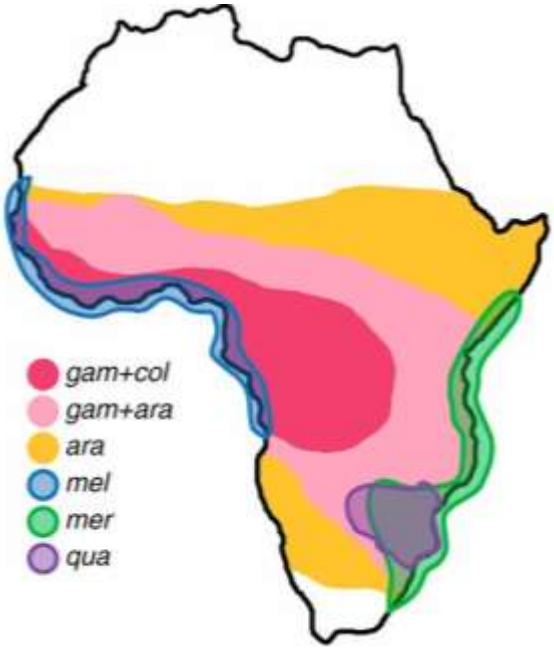


Figure 1.8. Distribution of six *Anopheles gambiae* complex members across Africa. Distributions and overlaps of *A. gambiae* (gam), *A. coluzzii* (col), *A. arabiensis* (ara), *A. melas* (mel), *A. merus* (mer) and *A. quadrianulatus* (qua) ranges. Adapted from Fontaine *et al*., 2015.

Beyond *A. gambiae* and *A. coluzzii*, other closely related anophelines are also often found with overlapping ranges (Figure 1.8). F1 hybrid males of these recently diverged crosses are usually sterile (Davidson, 1964), the heterogametic sex in accordance with Haldane's rule – (Haldane 1922; Schilthuizen, Giesbers and Beukeboom, 2011). However, recent research has revealed that divergence with gene flow is a trait found more widely across *Anopheles*, suggesting fertile hybrid females are an effective conduit for gene flow between species (Fontaine *et al*., 2014). Research into *A.gambiae* x *A. arabiensis* hybridisation found that, in addition to male sterility, recessive factors on the *A. gambiae* X were incompatible with factors on *A. arabiensis* autosomes (Slotman, della Torre and Powell, 2005) but contemporary gene flow has been shown between these species, with hybrid generations beyond F1 and much back crossing detected (Weetman *et al*., 2014). *Anopheles* are, therefore, valuable models for SGF. Populations and species with varying levels of divergence and isolation allow the possibility of finding speciation genes before they are lost (Via, 2009; Turner and Hahn, 2010; Weetman et al., 2012). Genomic research into

Anopheles gambiae and *coluzzii* is expedited through excellent data infrastructure, including whole genome reference sequences being available for both generalised and the specific forms (the sister species) of the insect (Holt *et al*., 2002; Lawniczak *et al*., 2010) and the powerful online bioinformatics resource, VectorBase (Lawson *et al*., 2009).

1.10 Insecticide resistance

As vector control campaigns generate strong anthropogenic changes to ecology with concomitant selective pressures on their targets, malaria mosquitoes also present an excellent system for studying the ecological adaptation of insecticide resistance. Resistance in mosquitoes is increasing (Ranson *et al*., 2001), has been shown to evolve over short time scales once populations are exposed to insecticide vector control campaigns and poses a real threat to the efficacy of these campaigns (Chandre *et al*., 1999; Brooke *et al*., 2001; World Health Organisation, 2012). Recent research has quantified the importance of insecticides (and therefore the threat of resistance) by the scale of their impact to malaria control; through evaluation of malaria surveys and control campaigns, Bhatt *et al*. show the great successes in the battle against malaria (~663 million cases prevented since 2015) have been overwhelmingly driven by ITN campaigns (68% of averted cases). In concert with other research reporting resistance in 27 Sub-Saharan countries (WHO, 2013) and revealing ITNs provide little protection where mosquitoes display resistance (due to every day wear and tear introducing holes in nets) (Asidi *et al*., 2012), there is danger of losing ground to the parasite and therefore pressure to better understand insecticide resistance and its underlying genetics.

Resistance to insecticides is generally split into two modes of action, target site mutations and metabolic resistance, though other forms such as behavioural and cuticular may also be important in malaria control (Ranson *et al*., 2011).The four chemical classes of insecticides (organophosphates (OPs), carbamates (CMs), Pyrethroids and organochlorines (such as DDT)) have only two targets, both within insect nervous system (Weill *et al*., 2004; Davies *et al*., 2007a). OPs and CMs affect acetylcholine esterase enzyme causing accumulation of the neurotransmitter acetylcholine (Čolović *et al*., 2013), while DDT and pyrethroids affect the voltage gated sodium channel (VGSC) by stopping it closing (Davies *et al*., 2007a). These proteins are found on the postsynaptic surface of neurons and binding of insecticides disrupts neurotransmission leading to hyperstimulation (a continuous action potential) and death (Davies *et al*., 2007a; Čolović *et al*., 2013). As these insecticides bind to and alter the function of proteins, the evolution of target site mutations which alter the confirmation of the

binding site and reduce the insecticide binding affinity can produce strong fitness effects where mosquitoes are exposed.

Well documented examples of fitness inducing target site mutations in malaria vectors include an amino acid change in the acetylcholine esterase 1 gene (*Ace-1-114S*) conferring resistance to CM and OP insecticides (Weill *et al*., 2004; Djogbénou *et al*., 2007; Oh *et al*., 2007). Though this mutation increases resistance (Weill *et al*., 2004), both the fitness cost of an altered neurotransmitter receptor (Djogbénou, Noel and Agnew, 2010) and recent research revealing that the mutation is often found in individuals carrying multiple copies of the Ace-1 gene (copies with and without the mutation) has led to suggestions that duplication allows retention of resistance while ameliorating the fitness costs (Weetman *et al*., 2015). Target site resistance to pyrethroids, used in ITNs, has also been detected in malaria vectors (Ranson *et al*., 2000). Three mutations of the VGSC gene have been associated with resistance in *Anopheles* mosquitoes, two affecting the same codon *Vgsc-1014F* and *S* (Martinez-Torres *et* al., 1998; Ranson *et al*., 2000) and one with an additive effect found only on haplotypes already carrying *Vgsc-1014F, Vgsc-1575Y* (Jones *et al*., 2012a). These protein altering changes are referred to in the literature as knock down resistance (*kdr*) mutations and are found throughout mosquito populations across Africa, *Vgsc-1014F* and *Vgsc-1575Y* mainly in the west of the continent and *Vgsc-1014S* in the east (Ranson *et al*,. 2011; Jones *et al*., 2012a). Though DDT and pyrethroid introduction is relatively recent (1940s and 1970s respectively – Davies *et al*, 2007a), widespread use of these insecticides to control many different pest species has led to strong selection for resistance and over 30 different resistance linked VGSC mutations have already been detected across Insecta (Rinkevich *et al*., 2013).

Metabolic resistance, rather than selection acting on the targets of insecticides driving resistance, involves an increase in the rate of breakdown or sequestration of the toxin within the insect and, as in detoxification in mammals, cytochrome P450s are a primary gene family involved in xenobiotic metabolism (Ranson *et al*., 2011). As regulation of gene transcription can be *cis* (physically near to the gene) or *trans* (distant from the gene), detecting this mode of resistance and developing diagnostic SNP markers can be more difficult than target site mutations, instead qPCR or microarrays are used to evaluate and correlate upregulation of detoxification gene transcription with resistant phenotypes before *in vitro* functional testing can confirm that the candidate proteins metabolise the insecticides being used (Ranson *et al*.,

2011). Strong candidates for metabolic resistance in *Anopheles* come from the CYP6 family, *Cyp6P3* and its orthologues have been found repeatedly upregulated in resistant *A. gambiae* (Muller *et al*., 2008) and *A. funestus* (Wondji *et al*., 2009) populations and the gene has been shown to metabolise pyrethroids (Muller *et al*., 2008). Other gene families involved in xenobiotic metabolism have also been implicated in resistance include the esterases and glutathione S-transferases (Hemingway, 2000).

Though target site and metabolic modes of resistance are well described, other modes of insecticide resistance are less well studied. As the main vector control techniques, ITN and IRS, involve the mosquito coming into contact with treated surfaces, insecticide absorption is through the cuticle (Ranson *et al*., 2011). An increased thickness of cuticle may therefore afford protection and two genes involved in the cuticle synthesis pathway have been found upregulated in resistant mosquitoes by microarray (Vontas *et al*., 2007; Awolola *et al*., 2009), however, more work is required to verify the role of these genes in resistance. Changes in mosquito behaviour that reduce contact with insecticides may also confer resistance (Liu *et al*., 2006). There has been suggestion that feeding behaviours have shifted from indoors (endophagy) to outdoors (exophagy) in response to vector control campaigns using insecticides indoors (Reddy *et al*,. 2011), however, more data is required to reveal if this is an adaptive response, if behavioural plasticity could explain the result and if a genetic basis can be found (Ranson *et al*., 2011).

With large effective population sizes predicted (Athrey *et al*., 2012), resistance alleles (*e.g. kdr* mutations) may be circulating in populations at low frequencies prior to insecticide exposure allowing fast evolutionary response (Karasov *et al*., 2010) and with the strength of selection such that resistance loci can breach species barriers and allele frequencies go from undetectable to almost fixation in under 10 years (Lynd *et al*., 2010). Clearly insecticide resistance mechanisms and evolutionary drivers are complex, but mosquitoes evolving over tractable time scales allows their study. Elucidation of speciation dynamics and ecological adaptation in *Anopheles* malaria vectors is especially germane, as understanding the speciation and insecticide resistance evolution of the insect vectoring the parasite responsible for the deaths of ~584,000 people in 2013 (World Health Organisation, 2014) can only improve control campaigns. Particularly as recent and more ancient introgression, including the transfer medically relevant insecticide resistance loci, has been shown to be a major feature of these vector's genomes (Weill, *et al*., 2000; Weetman *et al*., 2012; Fontaine *et al*., 2015).

1.11 Aims

This project aims to exploit the wealth of anopheline genomic resources publically available and the genome sequencing skills and capacity of the Wellcome Trust Sanger Institute, to investigate speciation and genomics of adaptation in these important species. WGS of individuals from Ghana will enable high resolution analyses of the porous species barrier still evident between *A. gambiae* and *A. coluzzii* in this region of low admixture (Weetman *et al.*, 2012). A genome-wide view of the extent of recent insecticide resistance locus introgression should be possible. Sampling and sequencing individuals from Guinea Bissau, a region with the highest documented levels of hybridisation and gene flow (Caputo *et al.*, 2008; Weetman *et al.*, 2012) will generate a window into the processes of divergence with gene flow at an earlier point on the speciation continuum. With the evolution of insecticide resistance playing such an important role in the adaptation of malaria vectors to anthropogenic pressures, WGS and the pan-African sampling of the *Anopheles gambiae* 1000 Genomes Project will be harnessed to investigate the history of a gene important in resistance. Origins of variants conferring insecticide resistance will be explored alongside how gene flow moves haplotypes containing them across Africa and between species. In some regions where another species from the complex, *A. arabiensis*, is found sympatric with *A. gambiae* it appears to have evolved insecticide resistance more recently than its sister species. With hybridisation and gene flow established there is the potential for adaptive introgression to be driving resistance (Mawejje *et al.*, 2013; Weetman *et al.*, 2014). The evolution of resistance in *A. arabiensis* will be investigated with the first pool-sequencing association study on *Anopheles* mosquitoes.

1.12 Project objectives

- Use WGS data to investigate the adaptive gene flow of medically relevant loci between the anophelines sibling species *A. gambiae* and *A. coluzzii*.
- Investigate the use of relative and absolute measure of divergence used in genome scans.
- Explore the genomic topography of divergence between *A. gambiae* and *A. coluzzii* when high levels of gene flow are found.

- Use WGS data to both explore the evolutionary history of loci known to be involved in the adaptation to insecticide use and to find new candidate loci.

Chapter 2 – adapted from the manuscript (Clarkson *et al*., 2014) published in Nature Communications - doi:10.1038/ncomms5248

Adaptive introgression eliminates a major genomic island of divergence but not reproductive isolation between *Anopheles gambiae* sibling species

Chris S. Clarkson[1]†*, David Weetman[1]†, John Essandoh[1,2], Alexander E. Yawson[2,3], Gareth Maslen[4], Magnus Manske[4], Stuart G. Field[5], Mark Webster[6], Tiago Antão[7], Bronwyn MacInnis[4], Dominic Kwiatkowski[4,7] and Martin J. Donnelly[1,4]

[1]Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, UK.

[2]Cape Coast Department of Entomology and Wildlife, School of Biological Science, University of Cape Coast, Ghana.

[3]Biotechnology and Nuclear Agriculture Research Institute, Ghana Atomic Energy Commission, P.O. Box LG 80, Legon, Accra, Ghana.

[4]Malaria Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1RQ, UK.

[5]Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, Colorado, 80523, USA.

[6]18a Church Lane, Hornsey, London, N8 7BU.

[7]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, United Kingdom.

†equal contributors

2.1 Abstract

Adaptive introgression can provide novel genetic variation to fuel rapid evolutionary responses, though may be counterbalanced by potential for detrimental disruption of the recipient genomic background. We examined the extent and impact of recent introgression of a strongly-selected insecticide resistance mutation (*Vgsc-1014F*) located within one of two exceptionally large genomic islands of divergence separating the *Anopheles gambiae* species pair. Transfer of the *Vgsc* mutation resulted in homogenization of the entire genomic island region (≈ 1.5% of the genome) between species. Despite this massive disruption, introgression is clearly adaptive with a dramatic rise in frequency of *Vgsc-1014F* and no discernible impact on subsequent reproductive isolation between species. Our results show (1) how resilience of genomes to massive introgression can permit rapid adaptive response to anthropogenic selection and (2) that even extreme prominence of genomic islands of divergence can be an unreliable indicator of importance in speciation.

2.2 Introduction

Anthropogenic habitat changes present a difficult evolutionary challenge for both intentionally and unintentionally targeted organisms, because of the speed at which they occur. Introgressive hybridization between incompletely reproductively isolated species provides a mechanism for the rapid acquisition of novel genetic variation which can accelerate adaptive evolution, and is of recognized importance for plants (Hedrick, 2013). However, only a few clear cases have been demonstrated in animals, for example, the transfer of rodenticide tolerance between mouse species (Song *et al.*, 2011) and of wing colour patterns among *Heliconius* butterflies (Kronforst *et al.*, 2013; Pardo-Diaz *et al.*, 2012). A major obstacle to adaptive introgression is the rate at which recombination can separate beneficial genetic variants within an introgressed fragment from the wider donor background, the disruptive effect of which on epistasis within the recipient species genome is likely to be deleterious (Hansen *et al.*, 2013). This may be exacerbated if introgressed adaptive variants are located in low recombination regions, because the hitchhiked portions of the donor species' genome will take longer to eliminate. Furthermore, because low recombination regions often exhibit elevated interspecific differentiation (Carneiro, Ferrand and Nachman, 2008; Noor and Bennett, 2009; Renaut *et al.*, 2013), disruption by potentially adaptive introgression may be particularly acute if divergent selection on variants in the region underpins differentiation. Finally, if species are very closely related and much of the interspecific divergence of their genomes represented in low recombination regions, this

detrimental effect of introgression might impact reproductive isolation directly. However, the association of differentiation with divergent selection is controversial.

Low recombination regions are prone to enhanced drift, recurrent linked selection and recurrent hitchhiking, which can generate similar patterns in the genome to those predicted under strong divergent selection (Charlesworth, Nordborg and Charlesworth, 1997; Cutter and Payseur, 2013; Noor and Bennett, 2009; Turner and Hahn, 2010). Although usually very difficult in wild populations, recent anthropogenic selection allowed us to investigate the extent and impact of adaptive introgression into a major 'genomic island' region postulated to be involved in divergent selection between the *Anopheles gambiae* species pair (Turner and Hahn, 2007; Turner, Hahn and Nuzhdin, 2005).

The M and S forms of *A. gambiae* are morphologically indistinguishable and were originally identified by fixed differences in ribosomal DNA near the centromere of the X chromosome (della Torre *et al*., 2001). Though recently elevated to species status as *A. coluzzii* (M form) and *A. gambiae sensu stricto* (S form) (Coetzee *et al*., 2013), for continuity with past work we retain the nomenclature of M and S, but discuss how our results bear upon this formal species definition. Divergence of M and S is thought to be driven by ecological niche separation of larval habitats (Lehmann and Diabaté, 2008). Differences in swarming locations have also been documented (Diabaté *et al*., 2009), and even in mixed swarms mating is usually assortative (Dabiré *et al*., 2013). However, M and S lack postzygotic isolation in the laboratory (Diabaté, Dabiré and Millogo, 2007) and hybrids are found occasionally in wild populations, although this frequency varies with country (della Torre, Tu and Petrarca, 2005).

Turner *et al.* identified two large regions of the genome toward the centromeres of chromosomes X and 2L that exhibit exceptional divergence between M and S forms (2005). This novel discovery provided evidence compatible with mosaic genome models of ecological speciation with gene flow (Wu 2001; Wu and Ting, 2004) and helped to spur the field of speciation. Such 'genomic islands of divergence' are hypothesized to arise via selection acting on a small number of physically-linked variants, and grow through hitchhiking of additional physically-linked adaptive and neutral loci (Cutter and Payseur, 2013; Feder and Nosil, 2010; Smadja, Galindo and Butlin, 2008; Turner, Hahn and Nuzhdin, 2005; Via and West, 2008). Moreover, although hybrids may be selected against (Lee *et al*.,

2013a), there is clear evidence for at least some contemporary gene flow extending beyond the F_1 generation throughout the range in which M and S co-occur (Lee *et al*., 2013a; Reidenbach *et* al., 2012; Weetman *et al*., 2012), a key assumption of mosaic genome models of ecological speciation (Noor and Bennett 2009; Wu and Ting, 2004). Nevertheless, discovery of additional areas of genomic divergence supported theoretical concerns that the 2L and X genomic islands might be unrelated to speciation (Neafsey *et al*., 2010; Weetman *et al*., 2010; White *et al*., 2010), their size arising via recurrent background selection and hitchhiking in areas of extremely low recombination (Charlesworth, Nordborg and Charlesworth, 1997; Noor and Bennett 2009). Resolution of these competing hypotheses has been hindered by the complexity of phenotypic differences between the species pair (Lehmann and Diabaté, 2008), which make laboratory studies very difficult. As a consequence, the importance of large genomic islands in the speciation process remains unclear.

Malaria-transmitting mosquitoes are subjected to massive insecticidal pressure, which drives selection for rapid development of resistance (Denholm, Devine and Williamson, 2002; Jones *et al*., 2012a; Lynd *et al*., 2010). Non-synonymous mutations in one of the two target sites for insecticides important in vector control, the voltage gated sodium channel (VGSC), are of particular significance. In *A. gambiae* the best known mutation, *Vgsc-L1014F*, confers knockdown resistance (*kdr*) to DDT and pyrethroids via a conformational alteration which reduces binding affinity of the insecticide (Davies *et al*., 2007a). In West Africa, *Vgsc-1014F* introgressed recently from S to M forms (Weetman *et al*., 2010; Weill *et al*., 2000) and has subsequently increased dramatically in frequency in M (Dabiré *et al*., 2009; Lynd *et al*., 2010), consistent with strong anthropogenic selection (Lynd *et al*., 2010). The VGSC is located within the large genomic island of divergence on chromosome arm 2L. Therefore adaptive introgression and selection of *Vgsc-1014F* will result in reduced interform divergence, but the extent and impact of this genomic disruption is unknown. In *A. gambiae* from southern Ghana, where M and S are broadly sympatric, we show that the entire 2L genomic island introgressed with apparently negligible impact on reproductive isolation during a period of rapid *Vgsc-1014F* increase, suggesting that it is neither critical to speciation nor maintained by strong divergent selection.

2.3 Methods

2.3.1 Collections

Adult female mosquitoes used subsequently for whole genome sequencing were collected by aspiration from southern Ghana during the summer of 2007. Six locations (Appendix 2.8.1) were sampled to yield five *A. gambiae* S form (individuals homozygous for the *Vgsc-1014F* mutation) and ten *A. gambiae* M form (five individuals homozygous for the wildtype *1014L* and five homozygous for the introgressed *Vgsc-1014F* mutation). *1014F* is close to fixation in *A. gambiae* S form populations in this region so no homozygote *1014L* individuals were available. Multiple collection locations were necessary due to a sequencing protocol that required a high yield of extracted DNA. Prior to extraction all samples were stored dry over silica. Data for M/S hybrid frequencies in southern Ghana were collated from collections we have described elsewhere (Essandoh *et al.*, 2013; Mitchell *et al.*, 2012; Weetman *et al.*, 2010; Yawson *et al.*, 2004; Yawson *et al.*, 2007).

2.3.2 DNA extraction and sequencing

After morphological identification as *A. gambiae s.l.*, DNA was extracted from dried whole bodies using the DNeasy extraction kit (Qiagen). Species (within *A. gambiae s.l.*) were identified using a standard PCR diagnostic assay (Scott, Brogdon and Collins, 1993), with subsequent identification of molecular forms using the SINE diagnostic method (Santolamazza *et al.*, 2008). As Ghanaian sampling was concerned with *kdr* introgression, these 15 individuals were also genotyped for the presence of the voltage gated sodium channel mutation *Vgsc-1014F* using a TaqMan assay (Bass *et al.*, 2007). DNA quantification was carried out using Quant-iT PicoGreen dsDNA fluorimetric assays (Invitrogen), taking the mean concentration from two technical replicates. Sample libraries were cloned with 200-300 bp inserts, and 76 bp paired-end sequencing was conducted using an Illumina High Seq 2000 by the Wellcome Trust Sanger Institute (European Nucleotide Archive: ERS012670-ERS012684). Reads were aligned to the AgamP3 reference genome (Holt *et al.*, 2002) using BWA (Li and Durbin *et al.*, 2009); 82-90% of read pairs per sample successfully aligned to the reference sequence, giving a median read depth per sample of 7-20x. Variant calling was conducted using SAMtools mpileup and BCFtools to produce a raw vcf file which was filtered using vcfutils.pl varFilter -D100 (maximum read depth = 100) (Li *et al.*, 2009). Non-biallelic SNP loci were removed using VCFtools (Danecek *et al.*, 2011). Sequenced individuals were re-checked for read depth, particularly in regions of interest, and molecular form was validated in the genome sequence data via presence/absence of the SINE insertion (Santolamazza *et al.*, 2008) using LookSeq (Manske and Kwiatkowski, 2009).

2.3.4 Statistical analyses

Genomic divergence between mosquito sample groups and pairwise nucleotide diversity within groups, were calculated for every SNP using VCFtools version 0.1.9.0 (Danecek *et al*., 2011) via Weir and Cockerham's (Weir and Cockerham, 1984) estimator of F_{ST} (--weir-pop-fst) and π (--site-pi), respectively. 'Fixed' differences in per-SNP F_{ST} between the S and M-wild type sample groups were identified and used (1) as ancestry informative markers (AIMs) to study localized recombination in the M-*kdr* group, and (2) to estimate the proportion of fixed differences (d_f) within non-overlapping 50 kb windows (Ellegren *et al*., 2012). This represents an arbitrary size that simply represents a balance between resolution and minimizing impacts of any SNP calling errors, and is not tailored to any specific detailed recombination map which is unavailable for *A. gambiae*. Plots of F_{ST}, d_f and π against chromosomal position were produced from means of windows using custom Perl scripts (https://github.com/cclarkson/thesis_chapter_2/blob/master /mean_Fst.pl) and visualised with the statistical software package R (R Development Core Team, 2011). A 100-SNP stepping window size was chosen to visualize F_{ST} because this has been shown to produce accurate estimates of Weir and Cockerham's F_{ST} from low sample sizes (Willing, Dreyer and Van Oosterhout, 2012).

To identify putative genomic islands of divergence we tested for exceptional values of d_f by simulating 100,000 Poisson distributions based on the actual number of windows and AIMs and applying a window specific threshold scaled according to the SNP frequency within the window. As a conservative threshold for identification of clustering within a window we applied a Bonferroni-corrected upper percentile limit for d_f from simulations: only observed d_f values exceeding this were considered significant. Adjacent significant windows were considered part of the same island; however, we considered islands as continuous if a non-significant window intervened between significant windows, but this exceeded the upper 0.8 percentile limit of simulated d_f. Kernel plots and associated skewness and kurtosis statistics (Bulmer, 1979; Cramer, 1997; Joanest and Gill, 1998) were used to study the density distributions of F_{ST} values on each chromosome and were calculated and plotted using R (R Development Core Team, 2011). Additional metrics to study variation in the pairwise polymorphic site-frequency spectrum between individuals within a sample and absolute divergence between samples were calculated as Tajima's D (Tajima, 1989) and D_{xy} (Wakeley, 1996), respectively using VCFtools 0.1.9.0 (Danecek *et al*., 2011) and custom Python code with visualisation in R (R Development Core Team, 2011), with a portion of values cross-checked with DNASP v5 (Librado and Rozas, 2009) to ensure correct script performance.

Sequence data was "haploidized" for estimation of these parameters (following Ellegren *et al*., 2012) by randomly assigning alleles at heterozygous positions. Spearman correlation coefficients between descriptive statistics (none of which were normally distributed) for islands were calculated using SPSS v20.

2.4 Results

2.4.1 Historical impact of *kdr* on islands of divergence

In Turner *et al*. (2005), two divergent islands are detected between M and S form of *A. gambiae* in Cameroon. One of the divergent islands is found proximate to the 2L centromere, a region containing the VGSC (Turner *et al*., 2005). However, as the authors did not genotype the samples for *kdr* mutations, divergence in this region could have been driven by M or S samples carrying selectively swept *kdr* carrying haplotypes, rather than speciation linked factors or neutral processes as suggested (Turner *et al*., 2005). We used several approaches to establish *kdr* status. The collector of the samples used in the 2005 study was contacted to discover if *kdr* genotyping was carried out. Samples collected at the same time as the study from the main sample site, Tiko (9 of 14 samples), were assayed for another study and these were found to not carry *kdr* mutations (F. Tripet pers. comm.). Tiko is geographically close to the other sample sites used: ~4km from Mutengene and ~20km from Buea.

The authors of the paper sequenced a part of VGSC exon 31 ("exon 17") for 46 samples including the 14 included in the main analysis and found no divergence within forms, just fixed differences between forms. With exon 31 <8kb from the *kdr* containing exon 19 and with little recombination expected in this low recombination region close to the centromere (Pinto *et al*., 2007; Carneiro, Ferrand and Nachman, 2008), it appears unlikely that a *kdr* containing, divergent haplotype, was sweeping through either population, particularly when no *kdr* was detected at the main collection site. In a more recent study of M and S divergence, Weetman *et al*. (2012) have also shown divergence at the 2L centromere in Guinea Bissau populations known to have no *kdr*, Though not conclusive, the evidence suggest that *kdr* was not present in the Turner *et al*. study (2005) and that even if it was driving the signal, divergence between M and S is found in this region in other populations in the absence of *kdr* mutations.

2.4.2 Extent and impact of kdr introgression

We sequenced the whole genomes of 15 wild-caught Ghanaian *A. gambiae* from three groups: S homozygous for the *Vgsc-1014F kdr* mutation; wild-type M which lack *kdr* (M-wt); and M homozygous for the *kdr* allele which introgressed from S (M-*kdr*). Comparison of M-wt and S form shows divergence across all chromosomes (Figure 2.1a) concordant with previous low density genome scans of Ghanaian M and S (Weetman *et al*., 2010; Weetman *et al*., 2012) and high density SNP genotyping of samples from Mali, Burkina Faso and Cameroon (Neafsey *et al*., 2010; Reidenbach *et al*., 2012). However, the two large islands near the centromeres of 2L and X identified originally (Turner, Hahn and Nuzhdin, 2005) are most prominent (Figure 2.1a). Comparisons between the groups of samples show that over 3Mb, representing approximately 1.5% of the genome, and apparently encompassing the entire 2L island of divergence, has introgressed between species. Consequently divergence between M-*kdr* and S forms in this region of the genome has been eradicated (Figure 2.1b), and in turn high, localised differentiation between M-*kdr* and M-wt created by introgression (Figure 2.1c). Beyond the 2L island the genomes of M-*kdr* and M-wt are minimally differentiated (Figure 2.1c), suggesting that either only the 2L island region introgressed from F_1 hybrids, or, perhaps more likely, that larger introgressed fragments have reduced in size through backcrossing and recombination within the M form.

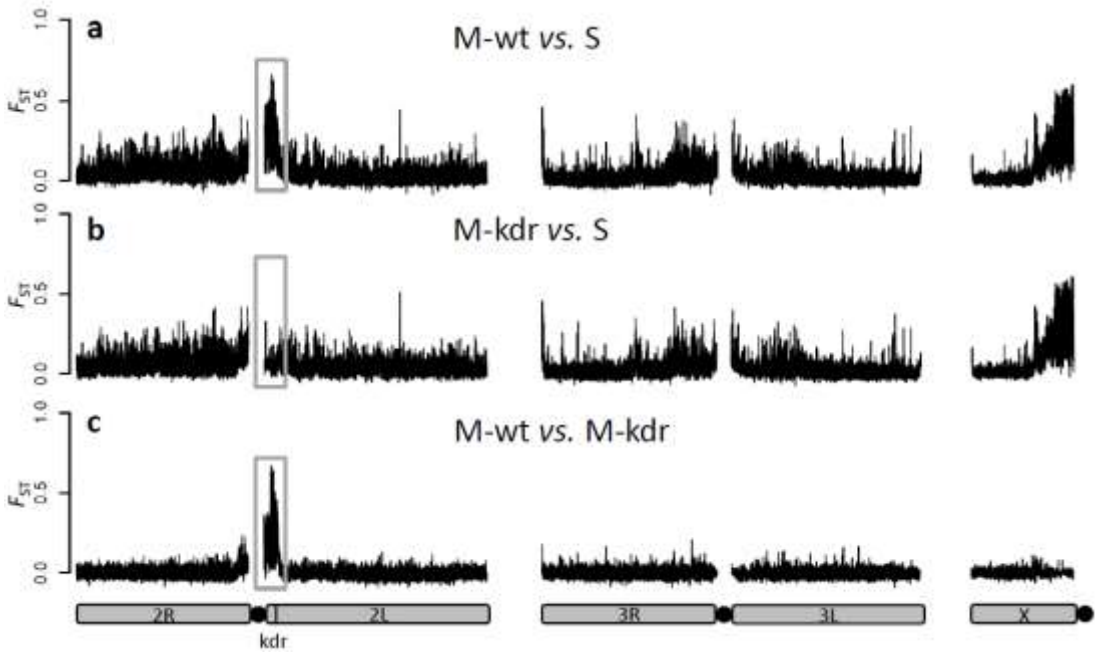


Figure 2.1. Manhattan plots showing F_{ST} –based pairwise divergence between groupings of *A. gambiae* S and M. Plots are based on mean F_{ST} in 100 SNP stepping windows for (**a**) M-wt *vs.* S, (**b**) M-*kdr vs.* S, (**c**) M-wt *vs.* M-*kdr*. Grey boxes highlight the 2L genomic island region involved in introgression. Chromosomes are shown by solid grey

bars and centromere positions by black circles. The position of the *kdr* (*Vgsc-1014F*) locus is shown on chromosome arm 2L.

We mapped the frequencies of M and S in larval collections from across southern Ghana (Figure 2.2). Overall, M and S were found at similar frequencies (55%: 45%), and though relative frequencies varied considerably among locations, M and S co-occurred in 15 of the 18 collection sites. In southern Ghanaian M forms, *Vgsc-1014F* is now present at consistently high frequency (mean ± s.d. = 0.79 ± 0.07; range = 0.67-0.90), in marked contrast to when first detected in 2002 (Figure 2.3a). This dramatic increase - to a frequency similar to that already present in S forms in 2002 (Yawson *et al*., 2004; Yawson *et al*., 2007) - is indicative of strong directional selection (Lynd *et al*., 2012). Despite the opportunity for hybridization afforded by widespread sympatry, frequencies of M/S hybrids throughout the period of dramatic *kdr* increase have remained low and stable (Figure 2.3b). This suggests (i) minimal impact of introgression of the 2L genomic island on reproductive isolation and (ii) that any divergent selection maintaining the island was much weaker than the directional selection driving the *kdr* mutation to high frequency. We examined whether a relatively low frequency of S forms in a collection site might limit opportunities for *kdr* introgression into M. However, there was no difference in current M-*kdr* frequencies between sites where S forms were rare (S frequency 0-0.09; *kdr* frequency = 0.78) and those where they were common (S frequency 0.48-0.96; *kdr* frequency = 0.82; t-test; t=0.82, P=0.41). Although relative frequencies of M and S in sampling locations may have varied during the period of *kdr* increase, available evidence of no current association between relative S frequencies and kdr frequency in the M form, points to relatively infrequent introgression of *kdr*, rather than manifold introgression events. In the following sections we examine evidence supporting hypotheses that might explain how introgression of such a large, highly divergent fragment could spread so rapidly and without apparent impact on reproductive isolation.

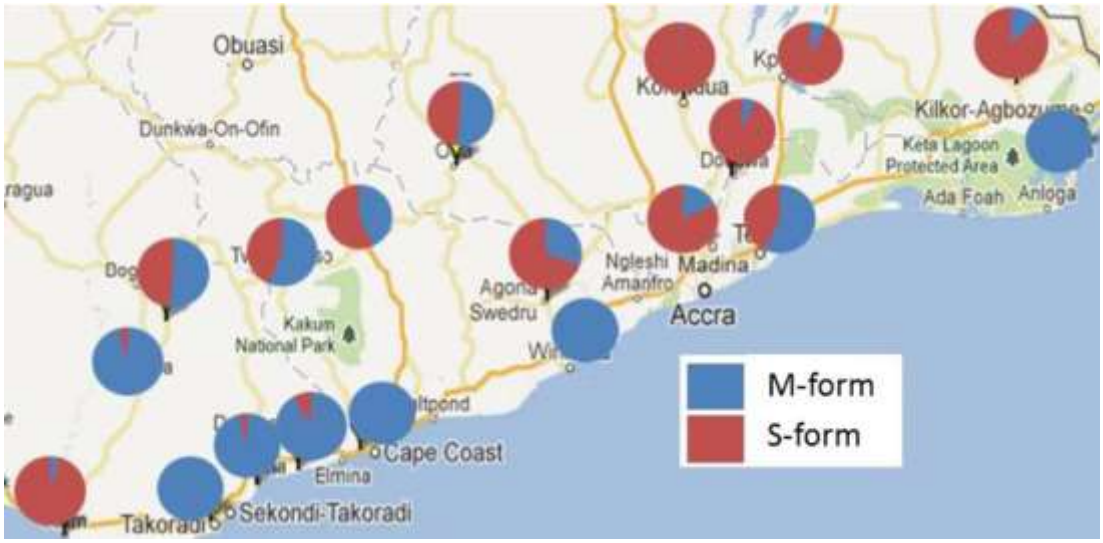


Figure 2.2. Distribution of the M and S forms of *A. gambiae* throughout southern Ghana. Pie charts show the relative frequency of each form (total N =846) in each site (N=18) detected in collections made in 2011.

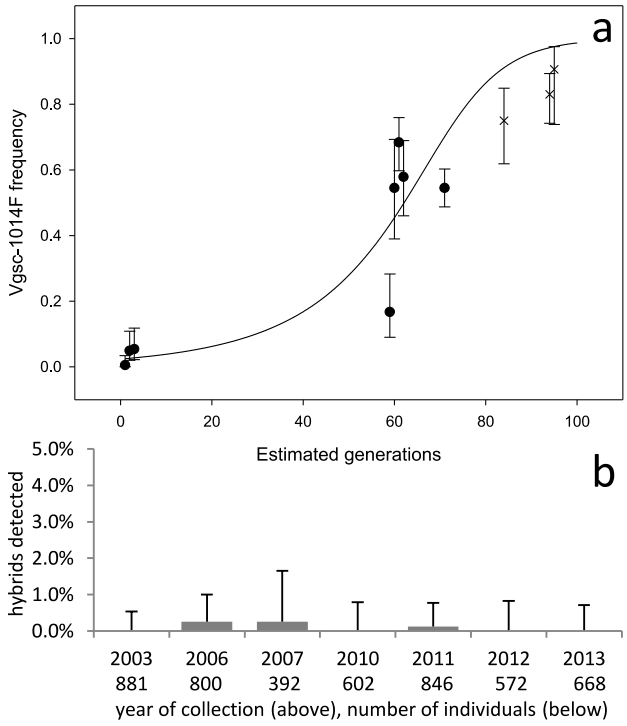


Figure 2.3. Spread of Vgsc-1014F kdr in M forms and M/S hybridization rates. (**a**) Increase in *Vgsc-1014F* frequency in M forms in Ghana: redrawn from ref. 33 (points shown as filled circles) with additional data points (points shown as x)). (**b**) Hybridization rates observed over a similar collection period with binomial 95% upper confidence intervals and sample size for each year.

2.4.3 Hypothesis 1

Only part of the 2L island introgressed, without key loci involved in reproductive isolation. Visual inspection of F_{ST}-based Manhattan plots (Figure 2.1) suggest that the entire 2L genomic island of divergence introgressed, but to examine this further we calculated mean pairwise nucleotide diversity (π) from the centromere across the first 5 Mb of the 2L chromosome arm (numbering on 2L starts at the centromere); a region exceeding the span of the genomic island. Neither S nor M-wt exhibited any evidence of reduced π (Figure 2.4a,b), though the S form does experience localized lower π relative to M, possibly due to the effects of the historical sweep around *Vgsc-1014F* in S (Lynd *et al*., 2012). In contrast, and as expected in a region currently undergoing a selective sweep, the M-*kdr* group shows a sharp drop in π (Figure 2.4c). However, unlike F_{ST} (Figure 2.1b,c), the signal from reduced nucleotide diversity does not span the entire 2L island (Figure 2.4c).

To investigate this disparity in more detail, we first identified ancestry informative loci (*i.e.* 'fixed' differences between the M-wt and S samples). Loci were then classified in each individual from the M-*kdr* sample as homozygous M-ancestry, homozygous S-ancestry or heterozygous (mixed ancestry) in the first 5Mb of the 2L chromosome arm (Figure 2.5). All M-*kdr* samples showed M-ancestry from approximately 3.3 to 5Mb onwards, and in three of the five M-*kdr* individuals, S-ancestry extended unbroken from approximately 3.3Mb back to the centromere. The other two M-*kdr* individuals showed near perfect mixed ancestry in the first 1.4Mb of the chromosome arm, with an identical transition point to homozygous S ancestry, indicating recombination at a single breakpoint within the 2L island (Figure 2.5). S-ancestry did not extend across the centromere into chromosome arm 2R in any M-*kdr* individual (results not shown). The integrity of the S island in eight out of the ten M-*kdr* chromosomes examined, the near 50:50 mixed ancestry in the other two from the centromere to the single shared breakpoint at 1.4Mb, suggests that recombination is recent. Thus introgression most likely did result in transfer the entire genomic island of divergence, which extends to 3.3 Mb, with recombination only just beginning to restore the M genomic background.

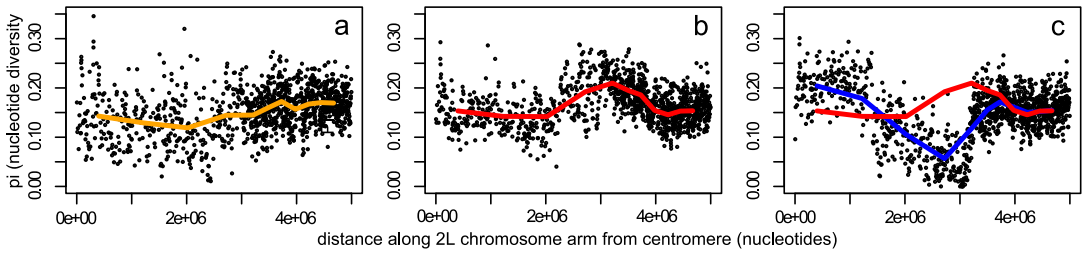


Figure 2.4. Nucleotide diversity (π) across the first 5 Mb of chromosome arm 2L, encompassing the genomic island region. For each sample group, individual points represent mean π in 100 bp stepping windows, whereas lines are smoothed by using a 10 kb stepping window scale in (a) S form, (b) M-wt and (c) M-*kdr*, represented by the blue line, with the M-wt line (red) included for comparison

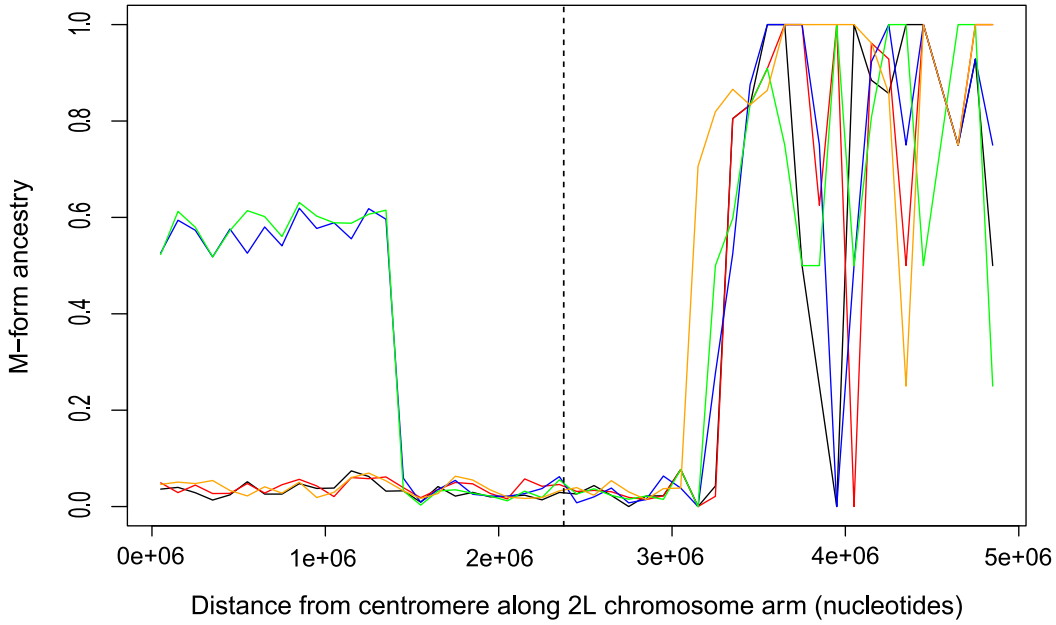


Figure 2.5. Analysis of recombination within the introgressed 2L genomic island. Lines show proportionate M form ancestry for each individual in the M-*kdr* group based on ancestry informative markers (fully diagnostic of M and S). The black dashed line indicates the location of the voltage gated sodium channel gene.

2.4.4 Hypothesis 2

The 2L island is selectively unimportant as speciation is advanced and divergence is genome-wide. Lack of impact of loss of the entire 2L genomic island might be because it merely represents the tip of a continuous distribution of divergence rather than a genomic island *per se*. To investigate this hypothesis we first examined the genomic distribution of

F_{ST}. In spite of the appearance of widespread, indeed potentially genome-wide, differentiation between M and S in the Manhattan plots (Figure 2.1a), inter-form differentiation is generally low, with a mean autosome-wide F_{ST} (± 95%CI) of only 0.032 ± 0.0002. Low genomic divergence, but high heterogeneity can be clearly seen from kernel density plots of the F_{ST} distributions for each chromosome arm (Figure 2.6) and the associated skew and kurtosis statistics (Appendix 2.8.2): all M-wt *vs.* S chromosome arm F_{ST} distributions are highly positively skewed and leptokurtic with long tails created by highly divergent SNPs (Figure 2.6).

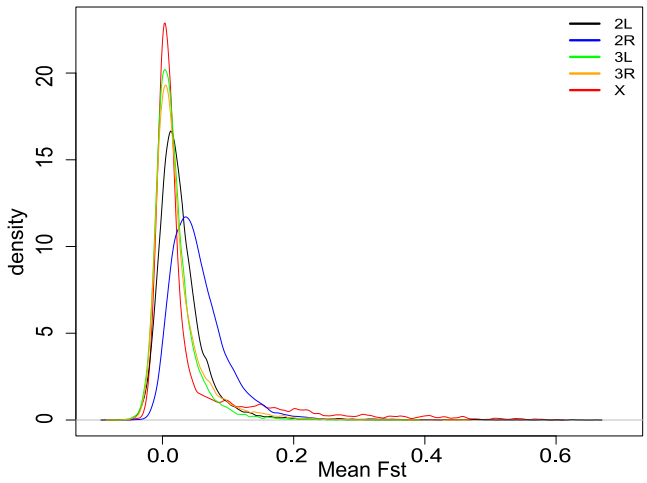


Figure 2.6. Kernel density plots of F_{ST} for M-wt. vs. S for each chromosome arm. F_{ST} was calculated using 100-SNP stepping windows. See Appendix 2.8.2 for associated mean, skew and kurtosis statistics.

To facilitate precise localization of areas of marked divergence (putative genomic islands) we utilized the ancestry informative loci, this time across the whole genome (0.24% of all 13,924,420 SNPs). From the proportion of fixed differences within 50kb windows (d_f) we defined non-contiguous windows significantly enriched for "ancestry informative loci" as distinct putative genomic islands of divergence. Plots of d_f suggest the presence of genomic islands (Figure 2.7) albeit highly variable in size and number across chromosome arms (Table 2.1). Over 80% of the putative islands are small, comprising of three or fewer adjacent significant 50kb windows (Table 2.1), whereas three were very large, the 2L island (3.3 Mb) and the two adjacent pericentromeric X islands (1.45 Mb and 4.9 Mb), which were likely merged in earlier low resolution analyses (Turner, Hahn and Nuzhdin, 2005; White *et al.*, 2010). Maximum and mean d_f were very strongly correlated with one another (Appendix 2.8.3) and with island size (Figure 2.8a, b), *i.e.* islands with higher d_f also tended to cover larger areas. Amongst islands both mean and maximum d_f were significantly positively

correlated with SNP frequency (Appendix 2.8.3), and though island size was not, it was notable that the largest islands had relatively few SNPs (Figure 2.8c), and also relatively few genes (Figure 2.8d). This contrast highlights the different patterns of polymorphism between smaller and very large islands, with the former exhibiting increasing SNP frequency with size, a relationship which breaks down for the largest islands.

Our results suggest that genomic divergence between M and S is both highly heterogeneous and largely restricted to islands. Moreover, the very large islands on 2L and X remain almost as prominent as originally suggested in early low resolution scanning (Turner, Hahn and Nuzhdin, 2005), and contain almost 45% of all significantly differentiated windows. In summary, though islands appear both far more numerous and thus cover more of the genome than originally thought (Turner, Hahn and Nuzhdin, 2005), divergence appears too heterogeneous in both island size and distribution to be considered as genomewide (Andrew and Rieseberg, 2013; Feder, Egan and Nosil, 2012); therefore hypothesis 2 is not supported.

Table 2.1. Size distribution of islands divergent between M and S.

size class (bp)	2L	2R	3L	3R	X	total	cumulative %
50000	10	29	7	16	2	64	55%
100000	1	10	1	3	3	18	70%
150000	2	6	1	4	0	13	81%
200000	0	2	0	2	0	4	85%
250000	0	4	0	0	0	4	88%
300000	0	3	0	2	0	5	92%
350000	0	3	0	0	0	3	95%
400000	0	0	0	1	0	1	96%
450000	0	0	0	0	0	0	96%
500000+	0	2	0	0	0	2	97%
1000000+	1	0	0	0	2	3	100%

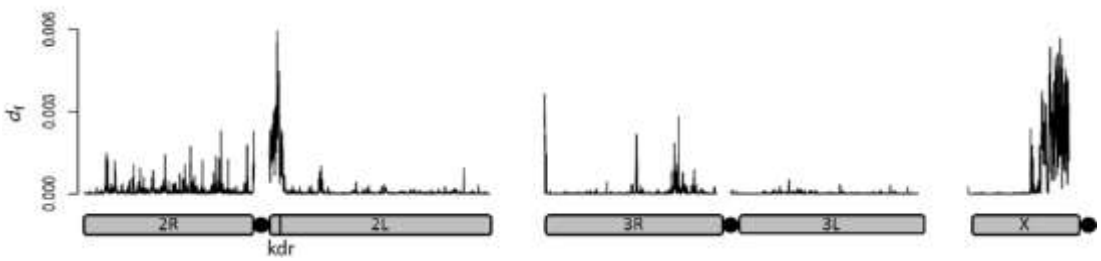


Figure 2.7. Genomic landscape of divergence between M and S. The y-axis shows the density of fixed differences between M-wt and S (d_f) in 50 kb stepping windows. Chromosomes and centromere position are shown by grey bars and black circles respectively; the position of the *kdr Vgsc-1014F* locus is shown.

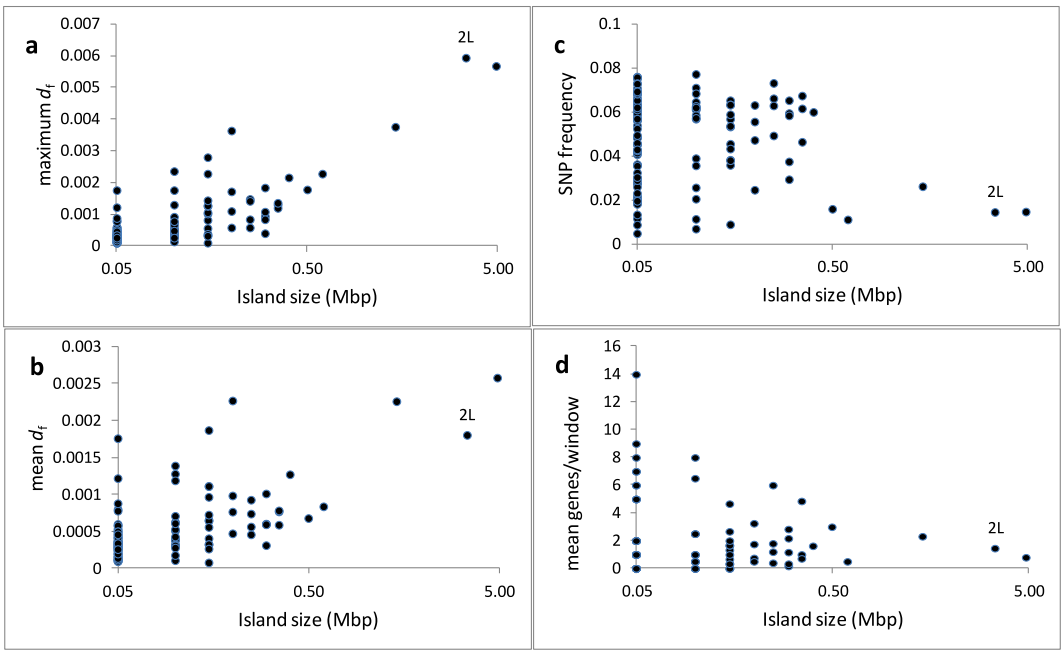


Figure 2.8. Scatterplots showing the relationships between the size of divergent genomic islands and descriptive statistics for diversity and differentiation. In each plot points are islands (total N=117). Owing to heterogeneity in island size a log scale is used in each plot; the major 2L genomic island indicated.

2.4.5 Hypothesis 3

Divergence of the 2L island results from processes reducing nucleotide diversity in low recombination regions rather than contemporary divergent selection. Our data suggest that genomic divergence between M and S is appropriately described by an island model, albeit

one involving many islands. Detailed recombination rate data are currently unavailable for the *A. gambiae* genome, but the location of the 2L island and the largest islands on the X chromosome near centromeres suggests that they are likely to experience reduced recombination (Carneiro, Ferrand and Nachman, 2008; Noor and Bennett, 2009; Pombi *et al.*, 2006; Stump *et al.*, 2005; Turner and Hahn, 2010), which is consistent with their relatively low gene and SNP densities. F_{ST} is inversely related to genetic diversity (estimated by number of segregating sites per window – Appendix 2.8.4), and strong differentiation could reflect the actions of forces, other than contemporary divergent selection, that reduce diversity which is then very slow to recover in low recombination regions (Cutter and Payseur, 2013; Noor and Bennett, 2009). Consequently we examined additional metrics for evidence of selection operating on islands, which might provide a means of partitioning historical signals of reduced diversity from recent divergent selection (Charlesworth, Nordborg and Charlesworth, 1997; Cutter and Payseur, 2013; Noor and Bennett, 2009). We first calculated D_{xy} (Takahata and Nei, 1985), a measure of absolute divergence of all nucleotide positions in a sequence, for 50 kb windows. Nevertheless, caution is required in application of D_{xy}, which is prone to high variance with smaller sample size, and is known to exhibit high stochastic variance among SNPs (Wakeley, 1996), both of which might affect genome scan analyses.

Consistent with, for example, the effects of background selection in low recombination regions (Charlesworth, Nordborg and Charlesworth, 1997; Noor and Bennett, 2009), D_{xy} was depressed near centromeres (Appendix 2.8.5) and peaks were not coincident with the islands identified using d_f (only one out of the 436 windows which comprise the islands exceeded a 0.99th percentile of D_{xy}). Such observations would appear to support hypothesis 3, that the islands could reflect historical rather than contemporary selective events (Noor and Bennett, 2009). However, D_{xy} was highly positively correlated with SNP frequency of islands (ρ = 0.89, P <<0.001) and also with its standard deviation within windows (ρ = 0.98, P <<0.001). Indeed, the relationship between the mean and standard deviation of D_{xy} is higher for islands generally (Figure 2.9), and the three very large islands (on X and 2L) show especially extreme relative standard deviation (Figure 2.9). In other words, the islands identified using d_f did contain large values of D_{xy}, but many monomorphic sites (i.e. a low SNP frequency) inevitably reduce values across a window, leading to exceptional variance for a given D_{xy} value. This will render D_{xy} extremely sensitive to the size of particular windows, with potential for ambiguous interpretation. Therefore, if covariates affecting D_{xy} are considered it could not usefully provide discrimination of competing hypotheses for our dataset. Moreover,

we note a high correlation (r=-0.5) between sequence depth and D_{xy}, suggesting sensitivity to sequencing error.

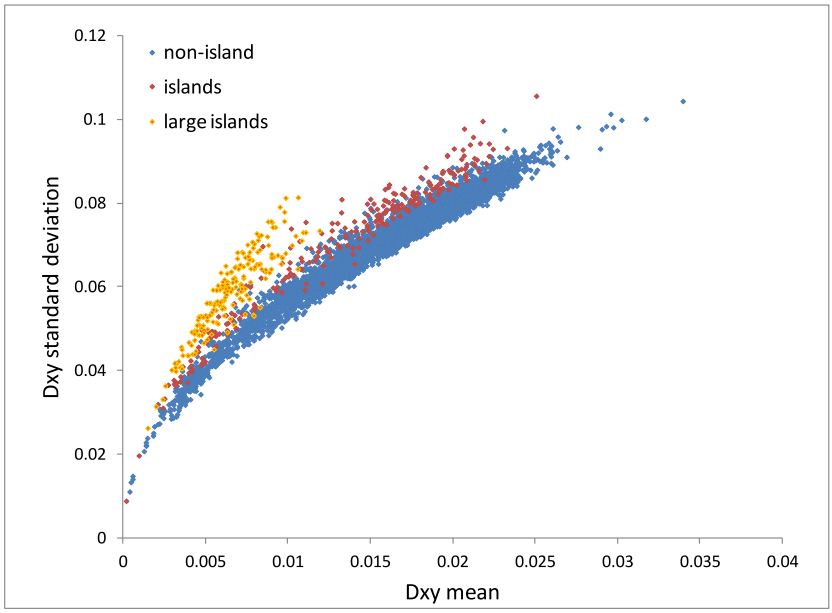


Figure 2.9. Scatterplot of absolute divergence, D_{xy}, plotted against its standard deviation. Points are means for every 50 kb window in the genome, with separate colours denoting windows from genomic islands of divergence and those in the three largest islands of divergence.

Secondly we calculated Tajima's D (Tajima, 1989), again for 50 kb windows; extreme values of D result from an imbalance between pairwise nucleotide diversity and the number of segregating sites, with negative values potentially indicating directional selection and positive values, balancing selection. In contrast to D_{xy} a low correlation between sequence depth and Tajima's D was found (r=0.19). There was no strong signal of directional selection around the 2L island in any group (Figure 2.10), with S forms actually exhibiting the highest positive values of Tajima's D within the 2L island, indicative of balancing selection (Appendix 2.8.6). This counterintuitive result seems unlikely to reflect balancing selection, but is concordant with theoretical expectations for a positive Tajima's D signal arising from a secondary selective sweep of the region (Chevin, Billiard and Hospital, 2008), which overlays an earlier sweep likely driven by *Vgsc-1014F*. The recently discovered resistance allele *Vgsc-1575Y* could provide a plausible candidate (Jones *et al.*, 2012a), as all the S form

individuals sequenced were N1575Y heterozygotes, although at around 3 Mb on 2L, the peak of Tajima's D is offset from the *Vgsc*.

Highly negative values of Tajima's D were almost entirely absent from the autosomes of M and S; though peaks were found throughout both of the two centromere-proximal X chromosome islands in M forms (Figure 2.10). In S, there was a notable clustering of negative Tajima's D peaks centred around 18.5 Mb, with extreme values (exceeding a two-tailed 99% threshold) just extending into the beginning of the largest X island (Figure 2.10) and thus clearly offset from peak interform divergence on X (Figures 2.1 and 2.7). Coincident outlying negative values were also found in the smaller island region within both M groups. Gene annotation term enrichment analysis identified a significant overrepresentation of genes linked with chitin synthesis genes in this region (Appendix 2.8.7 – due to table size this data can be accessed online - https://github.com/cclarkson/thesis_chapter_2/blob/master/Appendix_2.8.7). The negative Tajima's D signals, throughout the largest X islands in both M groups but not the S, suggest selection potentially acting within M forms rather than as divergent selection acting on both M and S as found in the smaller X island.

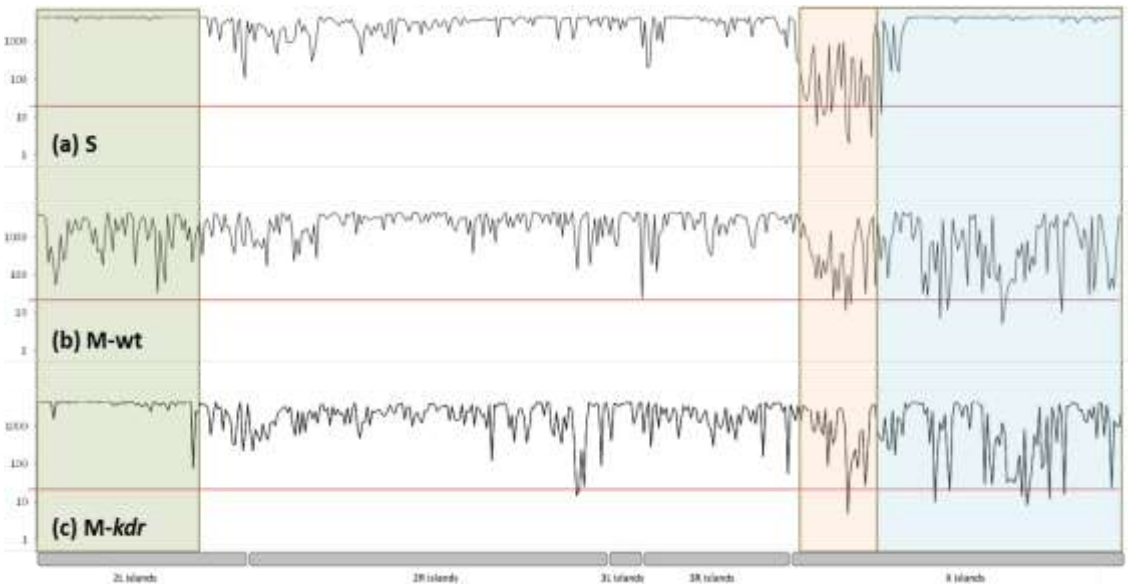


Figure 2.10. Evidence of directional selection from Tajima's D across all genomic islands in each group. Plots show ranks of Tajima's D (lower = more negative) for only the 438 significant windows comprising islands, arrayed in order of physical position. Ranks are calculated across all 4612 windows within each group. The red line shows the two-tailed

lower 99th percentile rank used as a threshold for extreme values. Windows within a major 2L island and in the pair of very large islands of divergence on the X chromosome are highlighted in shaded boxes.

The relationship between D_{xy} and its variance suggests unreliability because of extreme dependence on window size and a concerning potential for sensitivity to sequencing error. Tajima's D suggested some concordance of divergent regions with selection, though the likely conflation of signals within the 2L island region highlights that problems can occur with interpretation. However, Tajima's D yielded some signals of selection for each of the large X islands. In particular, the secondary island region on X, which lies outside of the pericentromeric heterochromatin region of extreme low recombination (Pombi *et al.*, 2006), is both significantly divergent *between* M and S and shows evidence of contemporary directional selection operating *within* M and S, supporting a novel hypothesis of involvement in ongoing divergence. In summary though hypothesis 3 is difficult to disprove conclusively, Tajima's D provided evidence of ongoing selection on the X islands (if not for the 2L island), and highlights the utility of combining multiple metrics in candidate region discovery.

2.5 Discussion

In this study we investigated a case of introgression between the most recently diverged species within the *A. gambiae* complex. The adaptive nature of the *Vgsc-1014F* mutation is clearly evident from its significant association with insecticide resistance (Jones *et al.*, 2012a) and its dramatic rate of increase in *A. gambiae* M forms. Introgression of *Vgsc1014F* from S to M forms has also been documented in Benin (Weill *et al.*, 2000), Cameroon (Etang *et al.*, 2009) and Burkina Faso (Dabiré *et al.*, 2009), with a similarly rapid increase in frequency observed in the latter. Moreover, though not explicitly considered by the authors, temporal variation in numbers of hybrids and backcrosses detected in a longitudinal study of a single village in Mali (Lee *et al.*, 2013a), might be linked to selection *for* introgression of *Vgsc-1014F* into M forms, rather than relaxed selection *against* hybrids and backcrosses linked to other, unknown environmental variations (Lee *et al.*, 2013a). The present data are unique in demonstrating the genomic extent of introgression, but studies of introgression from Mali and Cameroon, albeit based on only one or two SNPs in the 2L genomic island region outside of the *Vgsc* (Lee *et al.*, 2013a; Weetman *et al.*, 2012), and also the variety of locations from which *kdr* introgression has been recorded, suggests that our results are very unlikely to be restricted to southern Ghana.

Given the location of the *Vgsc* gene in the 2L pericentromeric region, which is thought to exhibit low recombination (Neafsey *et al*., 2010; Sharakhova *et al*., 2010), we hypothesized that a relatively extensive area might be affected by the selective sweep. In fact, the impacted area proved to be huge, exceeding 3 Mb (1.5% of the genome), and spanned the entirety of one of the two most prominent genomic islands of divergence between M and S. Coupled with hybridization data collected during the period of *Vgsc1014F* increase in M forms, this provided a natural test of whether the loss of a major genomic island of divergence reduces the reproductive isolation of M and S, as might be expected if the island contained genes critical to the speciation process; i.e. a 'speciation island' (Turner, Hahn and Nuzhdin, 2005). Our results do not support the designation of the 2L genomic island of divergence as a speciation island. M and S forms are extensively sympatric across southern Ghana, presenting widespread opportunity for hybridization. Yet hybridization rates appear stable throughout the period of rapid increase of introgressed *Vgsc1014F* to high frequency in M form populations across southern Ghana. It would appear that transfer of the entire island has had no discernible impact on reproductive isolation, allowing effective co-option of the adaptive *Vgsc-1014F* mutation into the M genomic background via adaptive introgression. The large pericentromeric speciation islands on separate chromosomes (X, 2L and 3L) are usually in strong linkage disequilibrium, which could imply epistatic selection (Turner and Hahn, 2010; White *et al*., 2010). If this were the case, it seems unlikely that the genome of M forms could tolerate such massive disruption without a major loss of fitness. By contrast our results suggest that any M form fitness cost is overcome by the increase in fitness from gaining the *Vgsc-1014F* mutation. It would seem therefore, that that any selective importance of the 2L island of divergence does not arise from its impact on reproductive isolation, and that it is not currently involved in speciation. Though some past involvement in divergence cannot be ruled out, our results highlight that large areas of inter-form divergence, however eye-catching, are not necessarily be under selective forces proportional to their size.

Reduced haplotypic diversity in the *Vgsc* of S-forms is evidence for recent selection (Lynd *et al*., 2010), which, prior to introgression of the *Vgsc-1014F* mutation and increase to high frequency in M forms, would have resulted in increased divergence. Although interpretation of the strong Tajima's D signal of selection on 2L was ambiguous, given low recombination in the 2L island, it is possible that a portion extending some way beyond the *Vgsc* might have been subjected to the sweep of *Vgsc-1014F*. This poses the question of whether M and

S divergence on 2L is simply a result of selection operating within S forms. Selection on *Vgsc-1014F* can be discounted as a general explanation because divergence in this region was first documented from comparison of M and S which both lacked the *Vgsc-1014F* mutations (Turner, Hahn and Nuzhdin, 2005). Selection within S alone is also not supported by comparative patterns of nucleotide diversity in the island region, levels of which are broadly similar in M-wt and S across the island region despite the historical selective sweeps in S (Jones *et al*., 2012a; Lynd *et al*., 2010) (Figure 2.4a,b). Apart from recent selection on *Vgsc-1014F*, is the 2L island under any contemporary directional selection at all or is its size an artefact of background selection? Unfortunately the additional metrics we applied (D_{xy} and Tajima's D) did not allow separation of these hypotheses but selection on *Vgsc-1014F* provides some additional insight. Given the very rapid increase in *Vgsc-1014F* frequency in M forms following introgression, any negative fitness consequences resulting from loss of alleles under selection within the 2L island must have been outweighed by insecticidal selection on *Vgsc-1014F*, for which we have estimated a selection coefficient of s=0.16 (Lynd *et al*., 2010). From the size of the introgressed fragment it is now apparent that this represents a net estimate for the 3.3 Mb 2L island of divergence, rather than *Vgsc-1014F* alone. Thus either selection on *Vgsc-1014F* is much stronger than initially estimated, or the total selection acting on all variants within the 2L island of divergence is weak.

If selection on such a large island appears weak, it is natural to question the importance of the other, often small, islands throughout the genome, and whether reduced recombination plays a key role in their formation (Renaut *et al*., 2013). SNP frequency data do not support the latter for many of the smaller islands, which, in contrast to the very large islands, often showed quite high densities of segregating sites, at odds with the expectation for a low recombination region. Nevertheless, differentiation of so many islands seems puzzling unless they are by-products of genome-wide divergence, which does not appear to fit their heterogeneity in size and distribution. To account for the genomically-widespread differences between M and S, when there is clear evidence for recent gene flow, Reidenbach *et al*. proposed an 'extrinsic' environmental hypothesis (2012). In this scenario hybridization occurs in infrequent bursts during unusual environmental conditions, with strong selection against introgressed individuals when typical conditions return. Recent data from a time series study in a Malian village appear consistent with this hypothesis (Lee *et al*., 2013a), though as noted above *Vgsc-1014F* introgression might also be involved. An alternative, and not mutually exclusive 'intrinsic' hypothesis, is that the nature of the genomic landscape of divergence provides permissiveness to gene flow. Selection dispersed across many islands is likely to be weak for the majority of loci, potentially enabling resilience to temporary loss of

(weakly-selected) islands until they can be restored by back-crossing. Searching for 'individual speciation genes' in such a landscape will thus be difficult because of low selection coefficients and frequent lack of correspondence between significant divergence and functionality.

Our results provide a natural 'loss of function' test of the 2L island, which bears many similarities in terms of likely recombination profile, polymorphism and divergence to the only other exceptionally large islands, located on the X chromosome. We think it is unwise to extrapolate from these commonalities a similar lack of importance for the X islands in speciation between M and S. X (or Z) chromosomes evolve relatively quickly owing to reduced effective population size and are known to be critically involved in development of reproductive isolating mechanisms between many species (Presgraves, 2008), including other less closely-related members of the *A. gambiae* species complex (Coluzzi *et al*., 2002; Slotman, della Torre and Powell, 2005). Moreover, and although proof of a speciation island must come from demonstrable function, Tajima's D, F_{ST} and d_f all provide signals consistent with selection acting around 17-19 Mb on the X chromosome in both M and S, and perhaps further toward the centromere in M forms. The signal of selection found in both M and S was primarily focused on the smaller of the two major X islands, the physical distance of which from the centromere may make it more likely that selection rather than just low recombination preserves its large size. Interestingly, whilst the largest island on X is always present when comparing M and S, this secondary X island area is absent from locales such as Guinea-Bissau, The Gambia and Senegal exhibiting exceptionally high hybridization rates (Nwakanma *et al.*, 2013; Weetman *et al.*, 2012). Follow-up studies on the role of this genomic region are warranted.

2.5.1 Conclusion

A multi-locus, resilient genomic architecture of divergence presents an interesting paradox for speciation theory. Typically the presence of substantial gene flow has been viewed as a signal of early stage incipient speciation (Wu and Ting, 2004), some way from the degree of reproductive isolation at which organisms might be recognized as 'good species'. However, it is becoming recognized now that gene flow between closely related 'good species' is extremely widespread (Feder, Egan and Nosil, 2012; Nosil, Funk and Ortiz-Barrientos, 2009). If selection is spread across numerous loci this may effectively provide intrinsic redundancy, and interspecific gene flow may actually be a long-lasting, stable state. In a

genomic landscape of generally weak but highly heterogeneous differentiation this state, though perhaps far from a highly differentiated 'endpoint' expected for species (Feder, Egan and Nosil, 2012), may be an important stage in genomic divergence, which can allow both adaptive introgression and protection of reproductive isolation. The molecular forms have recently been reclassified as *Anopheles gambiae s.s.* and *Anopheles coluzzii* (Coetzee *et al*., 2013), based primarily on evidence of reproductive isolation from relatively widespread genomic differentiation (Lawniczak *et al*., 2010; Neafsey *et al*., 2010; Reidenbach *et al*., 2012) and partial ecological niche partitioning (Constantini *et al*., 2009; Simard *et al*., 2009). Whilst we do not interpret our data as revealing truly genome-wide divergence, under either the 'extrinsic' or 'intrinsic' hypotheses outlined above, our results support this reclassification of M and S forms as species.

2.6 Acknowledgments

We thank Dr Robin Fencott for assistance with production of Perl scripts and three anonymous reviewers for comments and suggestions which significantly improved the manuscript. Sequencing, and genotyping support was provided by the Wellcome Trust Sanger Institute and the MalariaGEN resource centre. Additional funding support came from NIAID grant R01AI082734 (DW and MJD). CSC was supported by an LSTM Studentship, JE by Wellcome Trust MSc Fellowship in Public Health and Tropical Medicine WT094960MA and TA by a Sir Henry Wellcome Postdoctoral Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.7 Accession codes

The 15 *Anopheles gambiae* whole genome sequences have been deposited in the European Nucleotide Archive database under the accession codes ERS012670-ERS012684.

2.8 Appendix

Appendix 2.8.1 Sample information

Appendix 2.8.1. Samples used for whole genome sequencing.

origin	location	date	latitude	longitude	form	*Vgsc L1014F*
Ghana	Dawhenya	summer 2007	5.556	-0.19631	M	FF
Ghana	Okyereko	summer 2010	5.417	-0.600	M	FF
Ghana	Okyereko	summer 2010	5.417	-0.600	M	FF
Ghana	Okyereko	summer 2010	5.417	-0.600	M	FF
Ghana	Okyereko	summer 2010	5.417	-0.600	M	FF
Ghana	Dawhenya	summer 2007	5.556	-0.19631	M	LL
Ghana	Dawhenya	summer 2007	5.556	-0.19631	M	LL
Ghana	Kwamekyer	summer 2007	5.575	-0.6461	M	LL
Ghana	Kwamekyer	summer 2007	5.575	-0.6461	M	LL
Ghana	Kwamekyer	summer 2007	5.575	-0.6461	M	LL
Ghana	Awomberew	summer 2007	5.565	-0.6527	S	FF
Ghana	Awomberew	summer 2007	5.565	-0.6527	S	FF
Ghana	Gomoa, Onyadzi	summer 2007	5.359	-0.7061	S	FF
Ghana	Odumasi	summer 2007	5.907	-0.0833	S	FF
Ghana	Odumasi	summer 2007	5.907	-0.0833	S	FF

Appendix 2.8.2 Statistics

Appendix 2.8.2. Statistics associated with kernel plots of chromosomal differentiation

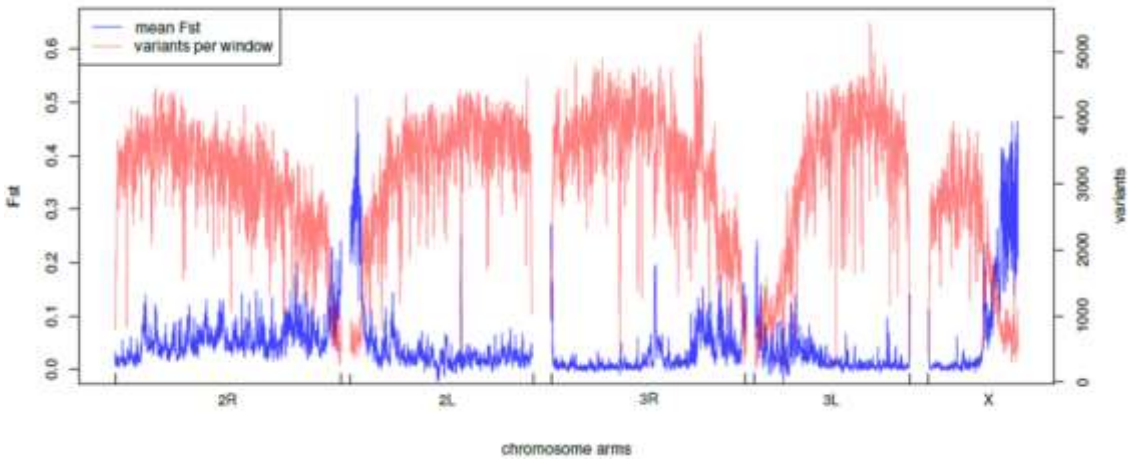
	2L	2R	3L	3R	X
Mean F_{ST}	0.0294	0.0545	0.0171	0.0225	0.0453
F_{ST} 95% CI ±	0.0006	0.0004	0.0004	0.0004	0.0018
Sample Skewness (G1)	4.362	1.367	2.440	2.495	2.870
Standard Error Skewness (SES)	0.014	0.013	0.015	0.013	0.024
Skew Zg1 (test stat) = G1/SES	316.9	106.5	158.3	192.3	118.6
Skew inference α=0.05	positive	positive	positive	positive	Positive
Level of skewness inference	high	high	high	high	High
Sample Excess Kurtosis (G2)	32.7	3.4	11.5	9.7	8.8
Standard Error Kurtosis (SEK)	0.028	0.026	0.031	0.026	0.048
Kurtosis Zg2 (test stat) = G2/SEK	1186.5	132.0	374.0	372.6	181.6
Excess Kurtosis Inference α=0.05	positive	positive	positive	positive	Positive
Direction	leptokurtic	leptokurtic	leptokurtic	leptokurtic	Leptokurtic

Appendix 2.8.3 Island statistics

Appendix 2.8.3 Relationships between island descriptive statistics

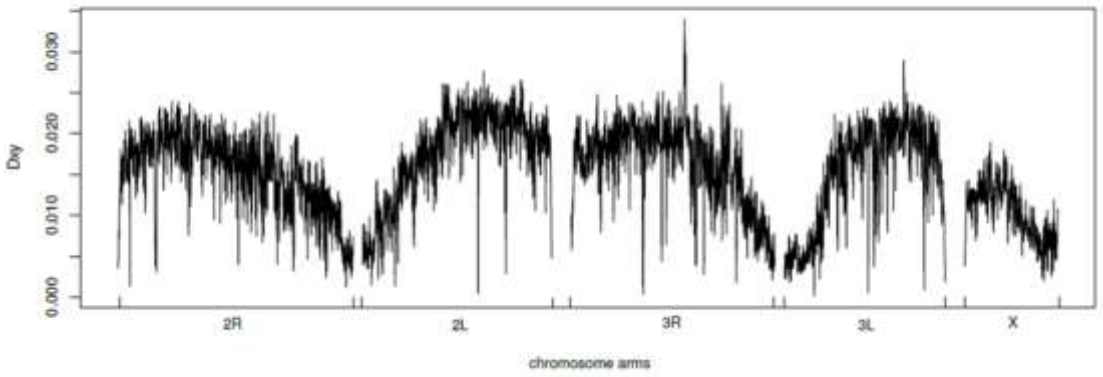
	size	d_f_mean	d_f_max	SNPs/window	genes/window
Size		<<0.001	<<0.001	0.823	0.648
d_f_mean	**0.564**		<<0.001	0.001	0.833
d_f_max	**0.656**	**0.983**		0.004	0.674
SNPs/window	-0.021	**0.300**	**0.266**		0.343
genes/window	0.043	-0.020	-0.039	-0.088	

Appendix 2.8.4 Divergence and variants per window



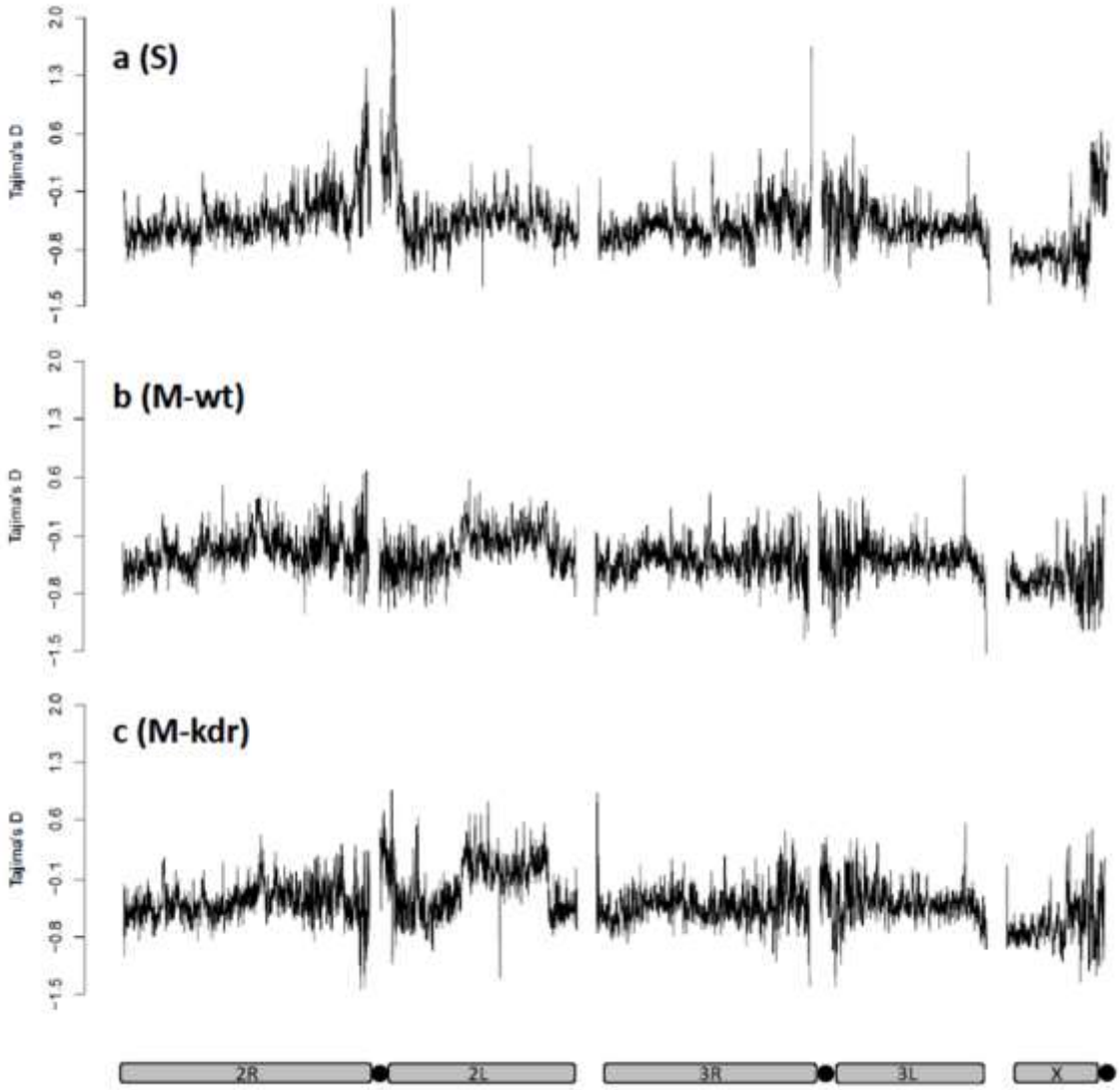
Appendix 2.8.4. F_{ST} –based pairwise divergence between M-wt and S *A. gambiae* with number of variants per window. The plot describes mean F_{ST} (blue) and the number of variants (red) in 100 SNP stepping windows across the whole genome (all chromosome arms) for M-wt *vs.* S.

Appendix 2.8.5 Genome-wide D_{xy}



Appendix 2.8.5. D_{xy} –based pairwise divergence between M-wt and S *A. gambiae*. D_{xy} calculated between M-wt and S *A. gambiae* across the genome (all chromosome arms) in 50kb stepping windows.

Appendix 2.8.6 Genome-wide Tajima's D



Appendix 2.8.6. Whole genome Tajima's D for the three groupings of *A. gambiae*. Plots are based on Tajima's D calculated in 50kb stepping windows for (**a**) S form, (**b**) M-wt and (**c**) M-*kdr*. Chromosomes are shown by solid grey bars and centromere positions by black circles.

Chapter 3

Species collapse in the "Wild West"? Genomic replacement by asymmetric introgression in an *Anopheles* hybrid zone

3.1 Abstract

Understanding speciation has been an enduring quest throughout the history of biological study and, with the advent of the genomic era's massive data production, the debate has only become more intense. Perhaps nowhere is speciation more medically important and controversial than in the *Anopheles* malaria vector mosquitoes. *Anopheles gambiae* was once thought to be undergoing incipient speciation, with two molecular forms M and S. However, with analyses finding genome-wide divergence between the two forms across most of their sympatric range, the forms were elevated to species and renamed *A. coluzzii* and *A. gambiae*. In the far-west of the range however, the story was not so clear cut. In Guinea Bissau, for example, much higher gene flow between the species was documented. Here we augment a trans-Guinea Bissau microsatellite study with whole genome sequence data, to examine the relationship between these two vector species under high gene flow conditions. Microsatellites revealed that the coastal region of the country was a hybridisation hot spot and genomic analyses demonstrate a marked difference in the *A. gambiae* from this region compared to inland populations. Asymmetric introgression is shown to have replaced much of the *A. gambiae* genome with that of *A. coluzzii* with potential for the collapse of these two species or even the generation of a new species through homoploid hybrid speciation. We demonstrate the speciation case between *A. gambiae* and *A. coluzzii* is far from closed and that in high gene flow environments, traditional species delimiters and concepts may need revising.

3.2 Introduction

3.2.1 Speciation in *Anopheles gambiae*

Since the publication of Darwin's most famous tome (Darwin, 1859), the biological debate about speciation has been a lively one; in recent times it would seem nowhere more so than in discussion of the major malaria vector mosquito, *Anopheles gambiae*. Early molecular work suggested that there may be two partially reproductively isolated ecotypes, named M and S forms, which were undergoing incipient speciation (della Torre *et al.,* 2001). Initially it was thought these two *A. gambiae* ecotypes were karyotypic forms because large inversions on chromosome 2, thought to be involved in ecological adaption, exhibited heterozygote deficit potentially due to population substructure (the Wahlund effect) (della Torre *et al.,* 2001; Wahlund, 1928). Additional data from later studies revealed that the inversions were not unique to M and S, thus not diagnostic, but it was shown that these morphologically indistinguishable ecotypes could be defined by molecular markers on the X chromosome (Favia *et al.,* 2001; Barnes *et al.,* 2005). Molecular, rather than karyotypic, forms of *A. gambiae* were now undergoing incipient speciation.

The first genomic work on the M and S molecular forms of *A. gambiae*, cemented the species as a model speciation system. Turner, Hahn and Nuzhdin's seminal paper (2005), used a microarray-based technique to show that two large, highly divergent genomic regions, in long range linkage disequilibrium (LD) across different chromosomes, were found when the molecular forms were compared. The authors suggested that because these were found against a background of very low genomewide divergence and, as it was known that M x S hybrids are viable (Diabaté *et al*., 2007), the M and S forms had been caught early in their divergence. Later a third large divergent island was revealed on the 3R chromosome arm by White *et al*. (2010). The molecular forms were touted as a rare example of what was classically termed sympatric speciation (Turner, Hahn and Nuhdzin, 2005). In recognition of the variability in the grain of geographical separation involved in sympatric speciation, it is now more commonly defined within the heading of speciation-with-gene flow. Sympatric speciation, with no geographical barriers to gene flow, is therefore the most extreme example of speciation-with-gene flow (Nosil, 2008). Based on earlier theoretical work suggesting how divergence with gene flow could begin and progress through a mosaic genome with some regions being protected from gene flow (Wu, 2001; Wu and Ting, 2004), the divergent regions, or genomic islands as they have been termed, were thought to contain the genetic drivers of speciation (Turner, Hahn and Nuhdzin, 2005).

A commentary on Turner and colleagues' paper (2005), suggested that because the genomic islands lay near centromeres, regions known to exhibit reduced recombination, it may be difficult to distinguish whether divergence was an effect of selection on adaptive variants or increased drift (Butlin and Roper, 2005b). Less prominent divergent islands found in on the 2R chromosome arm some distance from the centromere could therefore be more promising signals (Turner, Hahn and Nuhdzin, 2005). However, Turner and Hahn found these islands to not be universal across different populations of M and S in a later study (2007). As research on the system continued, controversy grew and the consensus began to change. It was suggested that there may actually be little realised contemporary gene flow between the molecular forms across most of their sympatric range and speciation-without-gene flow was suggested as an alternative scenario, with ancestral polymorphism driving the divergent genomic topography (White *et al.*, 2010). More momentum gathered from both theoretical commentaries (Noor and Bennett, 2009; Turner and Hahn, 2010) and high resolution genomic studies, which revealed genome wide divergence (Lawniczak 2010; Reidenbach 2012). In the light of these data, it was suggested the ecotypes were no longer undergoing incipient speciation, and the molecular forms were elevated to specific status (Coetzee *et al.*, 2013). M form was renamed as *A. coluzzii*, in memory of the prominent Italian vector biologist Mario Coluzzii, and the S form kept the original nomenclature, *A. gambiae*.

Realised gene flow between the *A. gambiae* molecular forms was considered low and extensive studies found hybrid counts across most of the M/S sympatric range to be <1% with the suggestion that there was little hybridisation progression beyond F1 (Tripet *et al.*, 2001; della Torre, Tu and Petrarca, 2005, Simard *et al.*, 2009). These data fit the more recent ideas of the forms being reproductively isolated units, good species in a 'Mayrian' sense (Mayr, 1942). However, regions with much higher gene flow have been described. In the far-west of the *A. gambiae* range, surveys found hybridisation rates of 7% in Gambia and >20% in Guinea Bissau (Caputo *et al.*, 2008; Oliveira *et al.*, 2008). Weetman *et al*. (2012) also showed that even in these high gene flow areas, with evidence of ongoing gene flow beyond F1 (only >F1 admixed individuals were found), some *gambiae/coluzzii* divergence remained. These results suggested that the mosaic model may still hold (Wu, 2001; Wu and Ting, 2004), and that the *A. gambiae vs. A. coluzzii* divergence might not simply represent segregating ancestral polymorphism (White *et al*., 2010).

Low gene flow and high isolation force a retrospective aspect to speciation research, to look back through evolutionary time and attempt to determine drivers of divergence. Where divergence and only partial reproductive isolation are found with high gene flow, barriers to flow and drivers of divergence can be identified before being confounded by drift and selective mechanisms post-isolation (Via, 2009). It is not yet clear whether these high gene flow conditions seen between *A. gambiae* and *A.* coluzzii result from secondary contact of diverged species (speciation collapse/hybrid speciation) or the early stages of divergence in the face of gene flow (ancestral conditions), but this 'aberrant' far-west re-opens the investigation of speciation within the *Anopheles gambiae* system.

3.2.2 Gene flow in Guinea Bissau

To study the high gene flow landscape and elucidate speciation in Guinea Bissau, over 600 female *A. gambiae* and *coluzzii* were collected from eight sites in 2010. The East-West sampling transect crossed Guinea Bissau's three major biotypes; A coastal region characterised mainly by mixed flooded forests and croplands; a central region where large patches of evergreen forest are present and a northeastern inland region characterized by shrubland and open deciduous forest (Figure 3.i1). Each individual was identified to species using both SINE and rDNA markers (Favia *et al.*, 2001; Barnes *et al.*, 2005), then was typed at 19 microsatellites, ten on the chromosome 3 and nine on X (Vincente *et al*., unpublished).

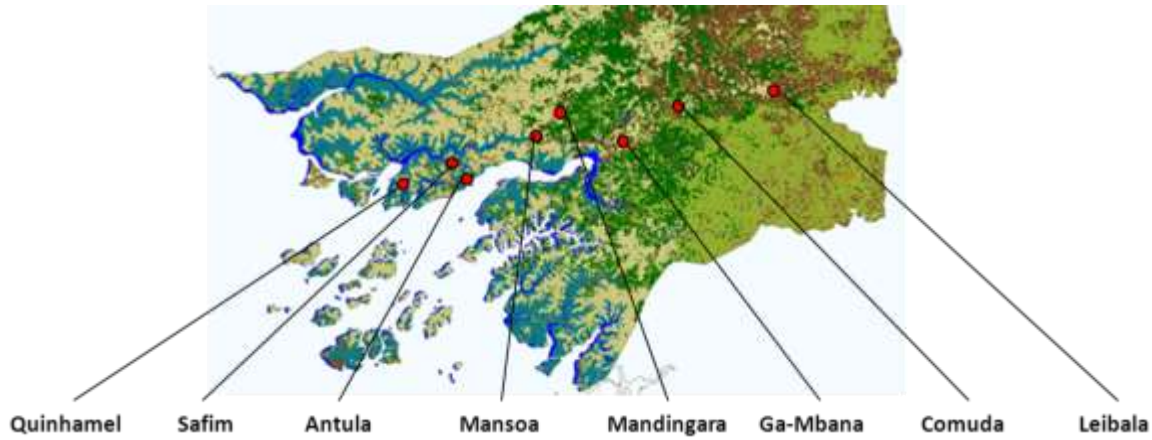


Figure 3.i1. Map of Guinea Bissau showing collection sites. (adapted from Vincente *et al*., unpublished)

Using this microsatellite data set, three genetic clusters were inferred using the Bayesian clustering algorithm in the software package Structure v2.3.3 (Prichard *et al*., 2000) (Figure 3.i2). These clusters loosely corresponded to the biotype zones across Guinea Bissau. Perhaps most striking was how species composition was driving the clustering. The inland cluster (cluster 2 - green) was composed almost entirely (91%) of *A. gambiae* and the central cluster (cluster 1 – red) was 88% *A. coluzzii* (Table i1). Neither of these clusters had >10% hybrids according to the two species markers on the X chromosome. The coastal cluster (3 – blue), however, was clearly driven by high numbers of hybrids (40%), higher even than had been previously reported (Oliveira *et al*., 2008). Thus even within Guinea Bissau, great variation in hybridisation rates were found, with the coastal area proving a hybridisation hotspot (Vincente *et al*., unpublished).

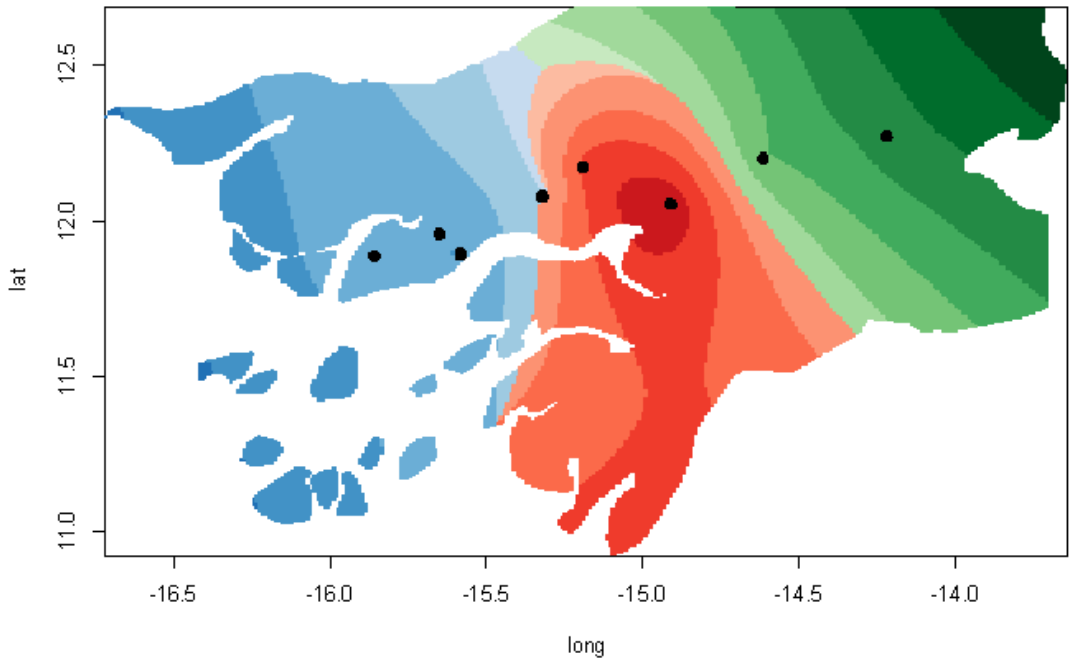


Figure 3.i2. Bayesian clustering. Map showing assignment probability densities for STRUCTURE analyses. Red: cluster 1, green: cluster 2, blue: cluster 3. (adapted from Vincente *et al.* unpublished)

Table 3.i1. Association between Bayesian genetic clusters (STRUCTURE) and molecular identification of species by IGS and SINE. Values represent the relative proportions of each species (determined by IGS/SINE markers) within each cluster. Highest proportions are highlighted in bold. *N*: total number of specimens assigned to each cluster (T_q = 0.5). (adapted from Vincente *et al.* unpublished)

		CLUSTER 1	CLUSTER 2	CLUSTER 3
IGS/SINE	*A. coluzzii*	**0.882**	0.022	0.022
	Hybrids	0.062	0.062	0.398
	A. gambiae	0.057	**0.916**	**0.580**
	N	211	178	231

A. gambiae and *A. coluzzii* are known to be ecological divergent, particularly with respect to larval habitat (reviewed in Lehman and Diabaté, 2008), therefore it is perhaps unsurprising that the differing central/inland biotypes result in differing species composition. However, the cause and effect (on the genomic scale) of the high hybridisation in the coastal region remains unknown. Here we augment the microsatellite dataset of Vincente *et al.* (unpublished) with whole genome sequencing data to test, at high resolution, the effects and direction of the high gene flow in the 'wild west' on the topography of divergence between these two malaria vectors.

3.3 Methods

3.3.1 *Anopheles gambiae* genome sequences

Guinea Bissau collections took place in October 2010. Mosquitoes were collected indoors overnight by CDC miniature light traps at all locations (Sudia and Chamberlain, 1962). Indoor resting collections with aspirators were also performed in Safim and Leibala. DNA was extracted according to Collins *et al*. (1987). Only mosquitoes which had been genotyped at both SINE (Barnes *et al.,* 2005) and rDNA (Favia *et al.,* 2001) markers as being homozygous *Anopheles gambiae* by PCR assay were selected for this study (Appendix 3.8.1). Samples were sequenced using Illumina technology by the Wellcome Trust Sanger Institute,

Cambridge. Four Ghanaian *A. gambiae* genomes were taken from the Clarkson *et al.* study from 2014 (Appendix 3.8.1). Collection sites, sequencing and variant calling information for these samples are contained within this previous publication (see also chapter 2).

3.3.2 SNP filtering and quality control

Samples came from different projects though all were sequenced by the Wellcome Trust Sanger Institute, Cambridge. To ensure that all samples were comparable and that there was confidence in the variants, we took the raw single nucleotide polymorphism (SNP) calls in variant call format (VCF) and conservatively filtered them using a custom Python script (https://github.com/cclarkson/thesis_chapter_3/blob/master/Antao_filter), rejecting individual's variants that displayed the following parameter values: GQ < 40, DP < 14, DP < median DP / 2 or DP > median DP x 2, MQ < 40, QD < 5, HRun > 3. The final two filtering rules are taken from the human 1000 Genomes methods (1000 Genomes Project Consortium, 2010), the additional rules were trialled against a presently-unpublished data set of *A. gambiae* laboratory crosses and were found to reduce Mendelian errors compared to the human filtering rules (pers. comm. T. Antão). It should also be noted that even the sample with lowest median DP has over three times that of the 4x human data (Appendix 3.8.1; 1000 Genomes Project Consortium, 2010). This sequencing depth, coupled with the conservative filtering, produced a data set with high confidence variant calls.

To further increase the quality of these data sets, no missingness was allowed. In analyses where just the three Guinea Bissau sites are compared, all filtered Guinea Bissau SNP calls were merged using *vcf-merge*, then positions missing data in any individual were removed using *vcftools --max-missing 1*, both tools from the VCFtools package (v0.1.12a) (Danecek *et al.*, 2011), to create a 'Guinea-Bissau-missingless' VCF file. To preserve the high resolution of analyses just using the more deeply sequenced Guinea Bissau samples, another data treatment was prepared to include the less deeply sequenced Ghanaian samples (Appendix 3.8.1, Appendix 3.8.2). For analyses including Ghana, these additional data were merged with the Guinea-Bissau-missingless VCF before missing sites were removed again, to produce an 'all-sample-missingless' set of variants. Though this data set was 66.2% smaller than Guinea-Bissau-missingless, 2,280,433 SNPs were still present across five chromosome arms (Appendix 3.8.2).

3.3.3 F_{ST} calculation

Pairwise population F_{ST} estimates were calculated using *vcftools --weir-fst-pop* (Danecek *et al.*, 2011). A windowed approach was taken, with mean F_{ST} calculated within non-overlapping 50kb windows. The window size used is arbitrary as no high resolution recombination map was available for *A. gambiae*; however 50kb allows ease of visualising signals of allele frequency divergence while minimising effects of non-biological artefacts such as sequencing errors. Windowed F_{ST} means were calculated using a custom Perl script (https://github.com /cclarkson/thesis_chapter_3/blob/master/Windowed_Fst.pl), chromosome arm means, standard errors and 95% confidence intervals were calculated using R (R Development Core Team, 2011). Manhattan plots were made with the R package qqman (Turner, 2014).

3.3.4 Ancestry informative markers

To investigate genome-wide introgression within Guinea Bissau *A. gambiae* samples, 733 ancestry informative markers (AIMs) were obtained from a study which used a 400k SNP chip to characterise divergence between *A. gambiae* and *A. coluzzii* from Cameroon and Mali (Neafsey *et al.*, 2010). SNPs which had allele frequency differences of ≥ 0.9 between the species were selected as being ancestrally informative. When applied to all Guinea Bissau variants that passed quality and missingness filters, 329 of the AIMs were present, 236 on the X chromosome and the remaining 93 distributed across the autosomes (Appendix 3.8.3). Percentage ancestry was assessed by scoring all Guinea Bissau individuals at each marker as homozygous *A. gambiae*, homozygous *A. coluzzii* or heterozygous (hybrid) ancestry. Autosomes and the X chromosome were examined separately in the percentage ancestry analysis as it has previously been demonstrated that X is less susceptible to introgression between species so may bias findings (Fontaine *et al.*, 2015), and because the X has a higher number of markers than all the autosomes combined due to the high and widespread species divergence found across it (Appendix 3.8.3; Clarkson *et al.*, 2014; Tuner, Hahn and Nuzhdin, 2005).

3.3.5 Principal component analysis

To produce an informative data set for principal component analysis (PCA), a minor allele frequency (MAF) cut-off was set to remove singletons. Singletons may be strong candidates for incorrect calls and the extremely high genetic diversity found in *A. gambiae* (Wilding *et*

al., 2009), as high as a SNP every two bases (Ag1000g Consortium, unpublished), meant that 26.3% of the data set was composed of singletons (Appendix 3.8.2), even after conservative filtering. High individual level differentiation (low MAF SNPs) may obscure the population level differences we are interested in detecting with PCA. LD pruning was also conducted to reduce SNP markers to those maximally independent and therefore maximally informative (Davis, Pandey and McKinney, 2011). Using the all-sample-missingless VCF and the software PLINK (v1.9) (Chang *et al.*, 2015), singletons were removed then data LD pruned using a 500 SNP window, sliding 100 SNPs at a time with an r^2 (linkage disequilibrium) threshold of 0.1 (Chang *et al.*, 2015).

PCA was then performed on the 3L and 3R chromosome arms using the smartpca function in the EIGENSOFT (6.0.1) package (Price *et al.*, 2006; Patterson *et al.*, 2006). Chromosome 3 was chosen for PCA as our interest lay in signals of gene flow and it is the least affected by confounding factors of adaption (Pombi *et al.*, 2008) and speciation (Fontaine *et al.*, 2015). 3L and 3R were qualitatively identical so here just 3L is shown (see Appendix 3.8.4 for 3R figure). VCFtools was used to calculate the mean depth for each individual (Danecek *et al.*, 2011).

3.4 Results

3.4.1 Guinea Bissau pairwise F_{ST}

Pairwise genome-wide comparisons of three Guinea Bissau populations of *A. gambiae* (previously *A. gambiae* S-form) using the fixation index, F_{ST}, revealed an unexpected topography. Despite all samples genotyping as *A. gambiae* at both the SINE (Barnes *et al.*, 2005) and rDNA loci (Favia *et al.*, 2001), only Safim *vs.* Antula, *i.e.* coastal *vs.* coastal (Figure 3.i1), shows the expected within-species Manhattan plot of low genomic divergence (Figure 3.1). Safim *vs.* Leibala and Leibala *vs.* Antula comparisons, both of which are coastal *vs.* inland (Figure 3.i1), reveal a signal of divergence that resembles that found when *A. gambiae* and *A. coluzzii* from Mali are compared (Figure 3.2).

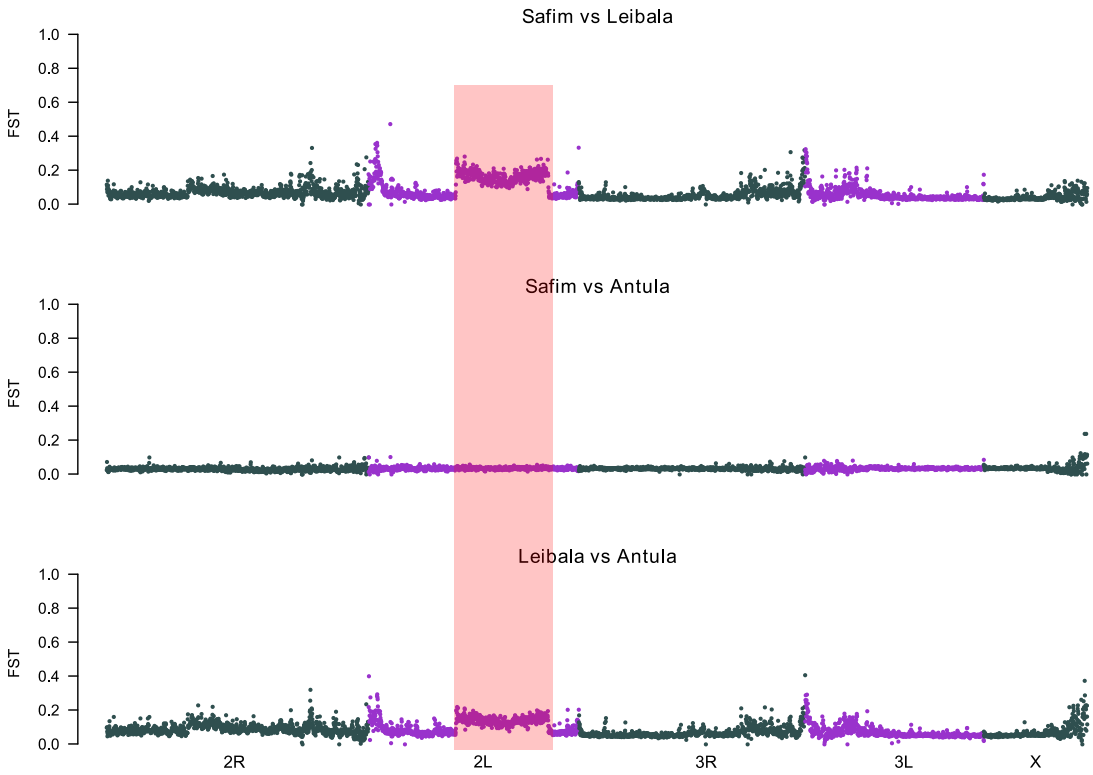


Figure 3.1. Mean pairwise F_{ST} between *A. gambiae* populations from Guinea Bissau. Plots show the three pairwise comparisons of genome-wide differentiation between Safim, Leibala and Antula sampling locations. Labels denote chromosome arm. A 50kb stepping window was used to generate the mean F_{ST}. Transparent red bar shows the location of the 2La inversion.

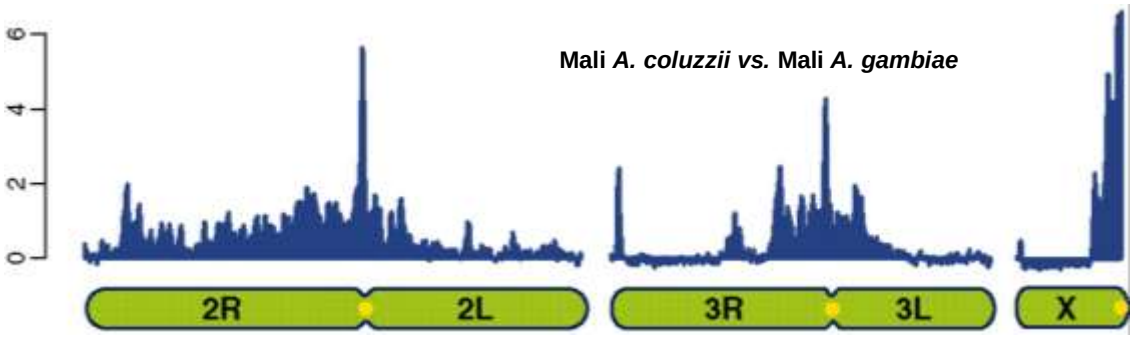


Figure 3.2. Malian *A. gambiae vs. A. coluzzii* whole genome divergence, representative of a typical pattern of differentiation between the species. Adapted from Neafsey *et al.*, 2010, Figure 1. Plot represents average difference in allelic intensity ratios measured over adjacent 50 SNP stepping windows. X axis represents chromosome arms, Y axis "relative local divergence (Z-score mode = 0)".

3.4.2. Ancestry informative markers

Species discriminating markers, developed from a study covering countries with more typical low gene flow, Cameroon and Mali (della Torre, Tu and Petrarca, 2005; Neafsey *et al*., 2010), allowed a genome wide exploration into the levels of introgression found in the exceptionally high gene flow regions of Guinea Bissau (Oliveira *et al*., 2008). Coastal *gambiae* samples (Antula and Safim) showed significantly higher numbers of autosomal introgressed *A. coluzzii* markers than the inland samples from Leibala (Figure 3.3a; Pearson's $X^2 = 333.13$, $df = 4$, $p = <0.0001$). *A. coluzzii* introgression was also higher on the X chromosome in coastal samples compared to inland (Figure 3.3b; Pearson's $X^2 = 88.57$, $df = 4$, $p = <0.0001$), although the percentage of introgressed makers was lower than found on autosomes (Heterogeneity $X^2 = 124.8$, $df = 4$, $p = <0.0001$).

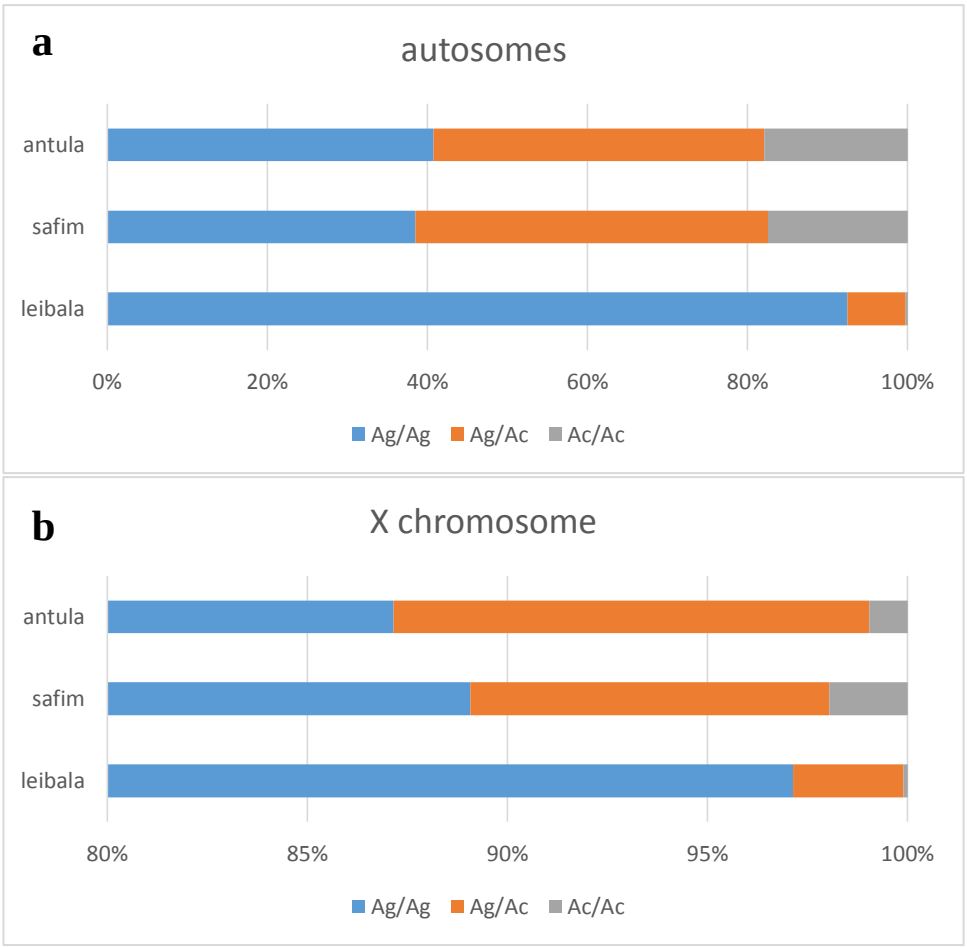


Figure 3.3. Percentage ancestry based on ancestry informative markers. Stacked bars show percentage of ancestry (Ag = A. *gambiae*, Ac = A. coluzzii) found in individuals sampled from three locations in Guinea Bissau. Percentage based on (**a**) 93 autosomal and (**b)** 236 X chromosomal markers, each scored as being homozygous for either one of the species or being of heterozygous ancestry (as an F1 hybrid would look). X axis differ between plots.

Lower levels of introgression on the X chromosome may be driven in part by a large stretch of 'pure' *A. gambiae* markers found in all individuals (Figure 3.4). This region, >3Mb (>4Mb but for one polymorphic marker) (Figure 3.4), homozygous for *A. gambiae* markers also revealed why, despite high levels of introgressed *A. coluzzii* genome, they still genotyped as *A. gambiae*: both SINE and rDNA markers (SINE and rDNA) lie within this peri-centromeric region on the X (Barnes *et al.*, 2005; Favia *et al.*, 2001). Large parts of the anopheline X chromosome have previously been noted as being resistant to gene flow between species; however the region we find homozygous *A. gambiae* is some distance from these (Fontaine *et al*., 2015). It should be noted however, that the AIMs used here do not cover these regions demonstrated previously to be gene flow resistant regions.

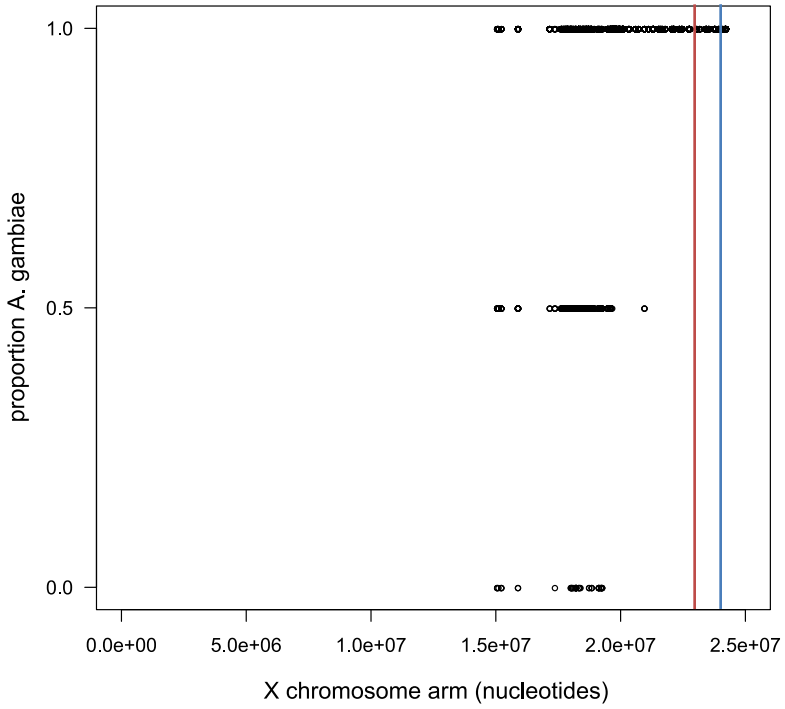


Figure 3.4. The distribution and proportion *A. gambiae* of ancestry informative markers across the X chromosome. 236 markers for each of the 21 Guinea Bissau individuals, genotyped as 0 = *A. coluzzii* homozygotes, 0.5 = *A.coluzzii/gambiae* heterozygotes or 1 = *A. gambiae* homozygotes. Vertical lines show approximate locations of species specific markers, red = SINE insertion, blue = rDNA polymorphism.

3.4.3 F_{ST} and variant density

Finding so many markers fixed for *A. gambiae* ancestry towards the X centromere raises doubts about the mean F_{ST} results in this region estimated using all SNPs (Figure 3.1). A higher resolution investigation into the X, using smaller (5 kb) windows, revealed in more detail the F_{ST} profile towards the centromere (Figure 3.5). The two coastal *vs.* inland comparisons (Safim *vs.* Leibala and Leibala *vs.* Antula), with relatively more divergent autosomes (Figure 3.1), demonstrated higher mean F_{ST} in windows towards X centromere than the geographically and ecologically close coastal *vs.* coastal comparison (Safim *vs.* Antula) (Figure 3.5). However, with 54 fixed *A. gambiae* AIMs in the last ~3Mb of all 21 Guinea Bissau individual's X chromosome, lower F_{ST} divergence was expected this region of the genome in all pairwise comparisons.

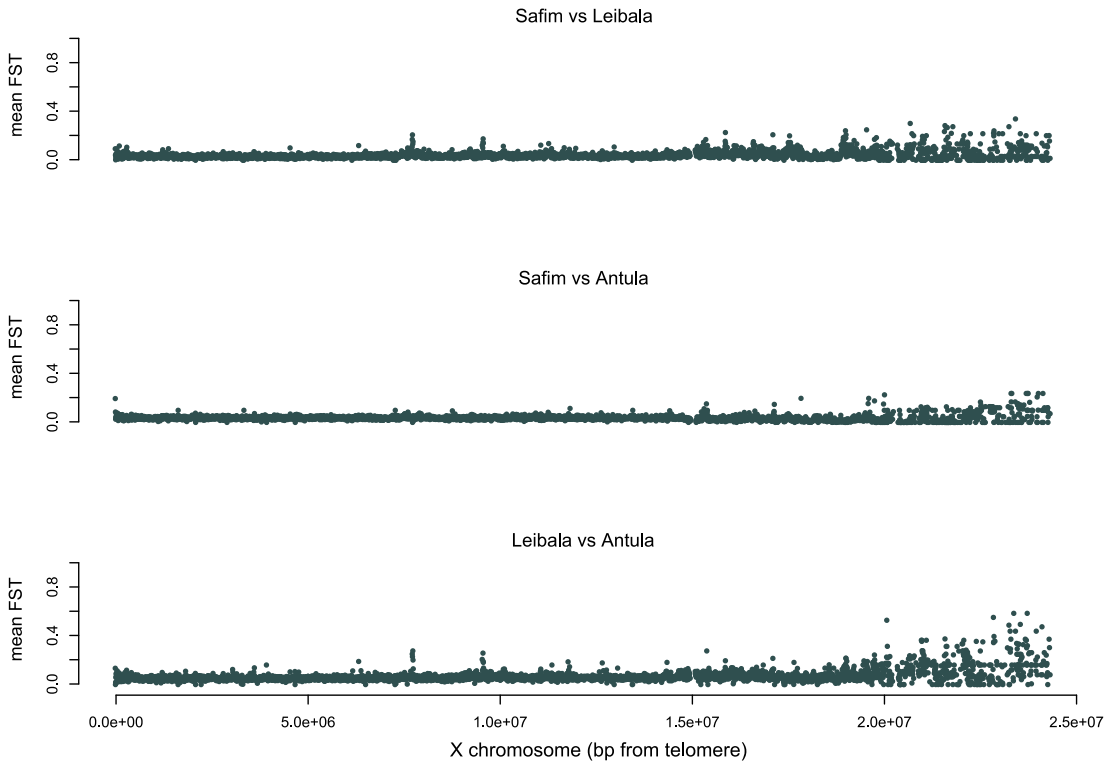


Figure 3.5. X chromosome mean pairwise F_{ST} between *A. gambiae* populations from Guinea Bissau. Plots show the three pairwise comparisons of differentiation between Safim, Leibala and Antula sampling locations. A 5kb stepping window was used to generate the mean F_{ST}.

Previous research into the use of relative measures of differentiation, such as F_{ST}, in regions of low recombination and low marker density regions (such as near to centromeres), has highlighted potential problems. Due to low variant density and a windowed approach, high variance in F_{ST} may be found, leading to spurious results driven by a small number of highly divergent SNPs in an otherwise low differentiation region (Cruickshank and Hahn, 2014). These factors, rather than biology, may be driving the divergent F_{ST} signal found pericentromeric on the X. Generally fewer SNPs are found across the X than on autosomes, but a striking drop in variant density was found in the last ~4Mb of the chromosome (Figure 3.6). The drop in SNP density coincided with an increase in the variance around mean F_{ST} in windows (Figure 3.7). Though AIM number is both high and fixed for *A. gambiae* ancestry in this region, a small number of highly diverged SNPs may be artificially inflating apparent differentiation. It should be noted that the X divergent island was caused by a wide spread of

windowed mean F_{ST} values, whereas other peaks see a concomitant rise in all windows compared to the surrounding regions (*e.g.* 2La region - Figure 3.1).

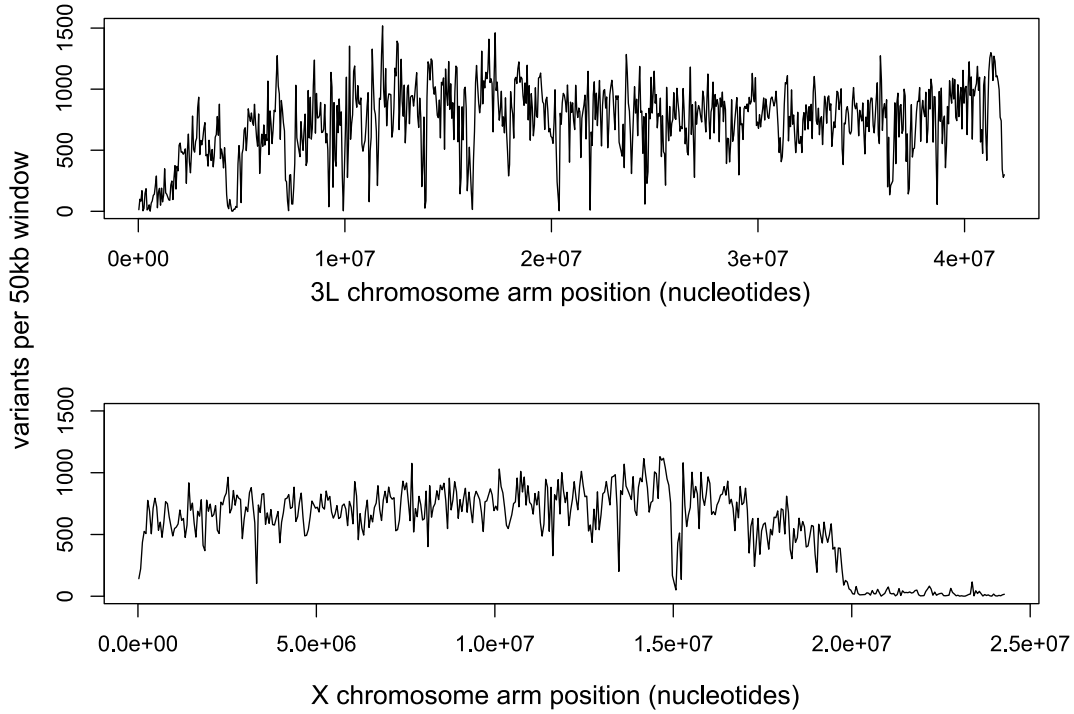


Figure 3.6. Number of variants within the 50kb windows used to calculate F_{ST} means across an autosome (3L) and the X chromosome.

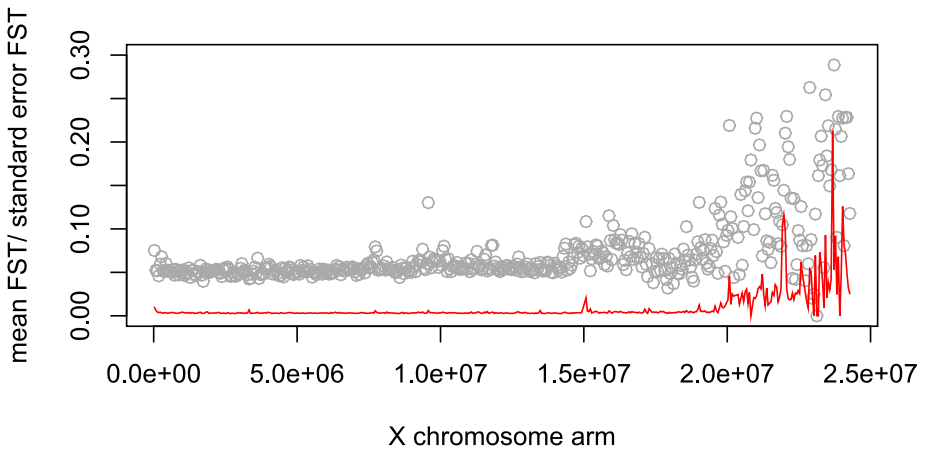


Figure 3.7. X chromosome pairwise mean F_{ST} and standard error between Leibala and Antula. For ease of standard error visualisation (red line), mean FST has been truncated at F_{ST}=0.3 (grey points). A 50kb stepping window was used to generate the mean F_{ST} and standard error.

3.4.4 Principal component analysis

The *A. gambiae* from Guinea Bissau were compared with samples of the same species from an African region with lower gene flow using PCA. Whole genome sequencing (WGS) samples were obtained from an earlier study in Ghana, a country where low levels of *A. gambiae* x *A. coluzzii* hybrids were found (reviewed in della Torre, Tu and Petrarca, 2005; Weetman *et al*., 2012). Eigenvalues revealed a dominance of principal component (PC) 1 (Figure 3.8), the relationship of it with PC2 and PC3 was investigated in detail.

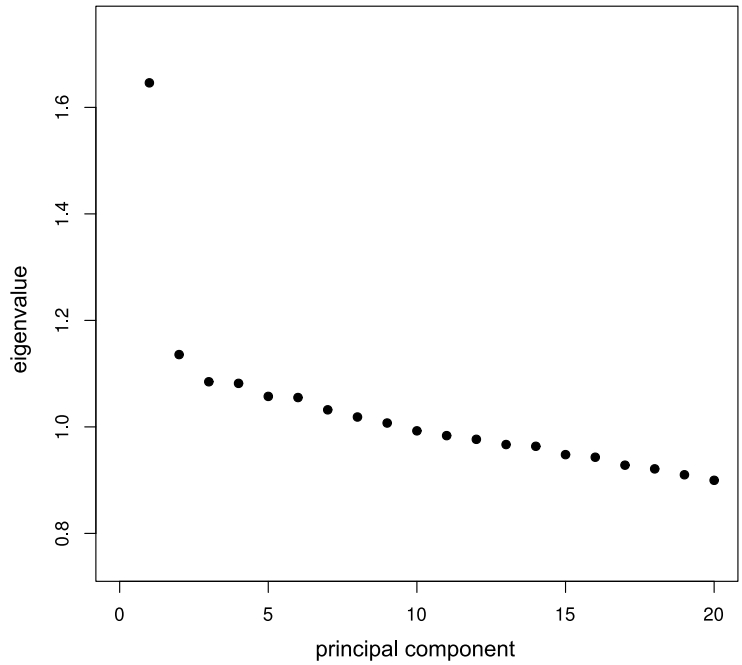


Figure 3.8. Relationship between principal component and eigenvalue for the 3L chromosome arm.

Principal component analysis with addition of *A. gambiae* samples from Ghana clustered Leibala (inland) samples with the Ghanaian samples (Figure 3.9a). All coastal Guinea Bissau samples, Safim and Antula, formed another cluster separated by PC1 (Figure 3.9a-b). These results, in concert with the F_{ST} and AIM analysis (Figure 3.1 and Figure 3.3), suggested that the Leibala samples are more similar to the *A. gambiae* found in countries with lower gene flow (Clarkson *et al*. 2014, Neafsey *et al*. 2010). However the Safim and Antula samples which genotyped as *A. gambiae* are actually autosomally *A. coluzzii*-like due to introgression. The species differences are highlighted on PC1 (Figure 3.9a-b), with PC2 and PC3 separating differences within the populations.

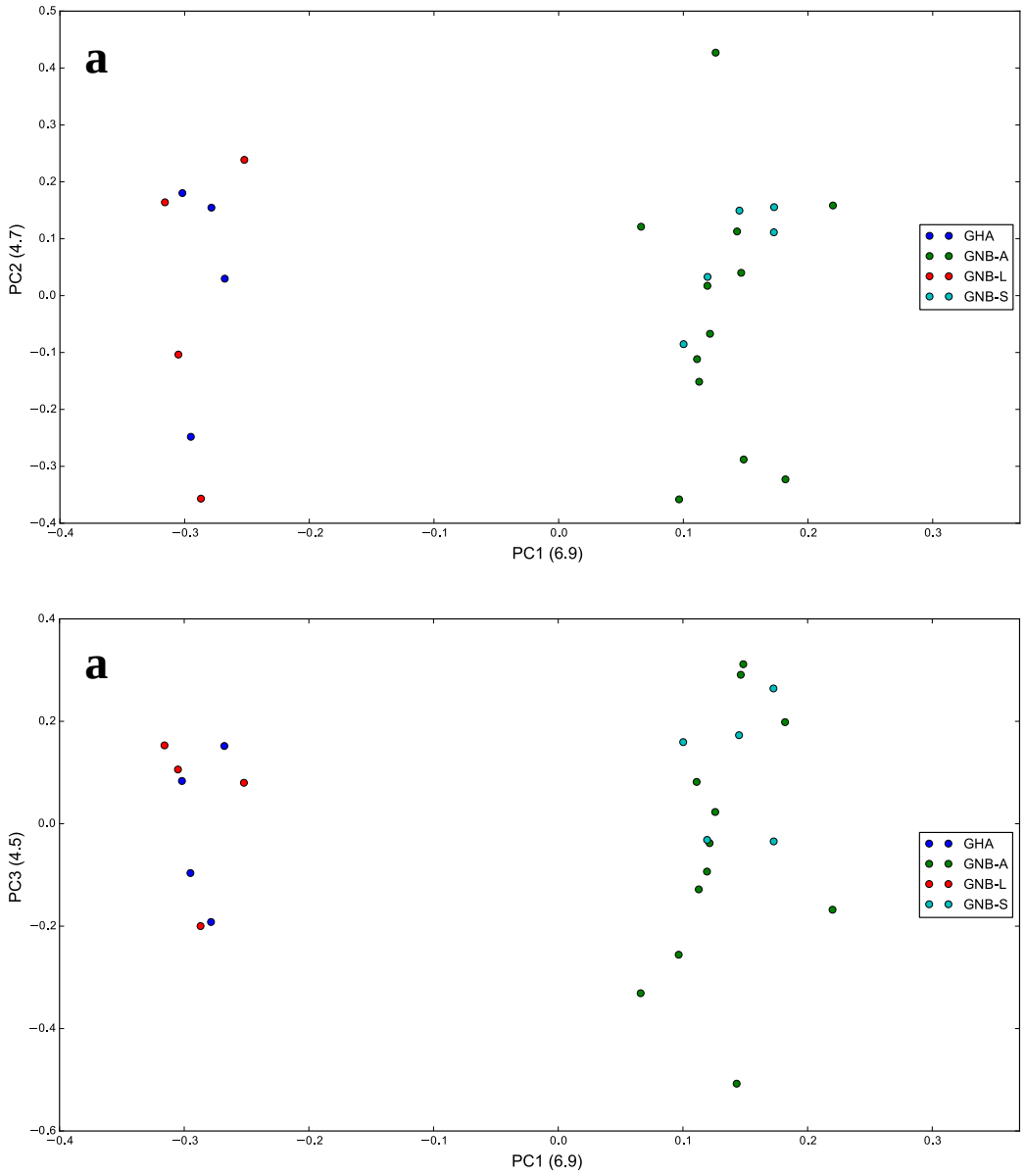


Figure 3.9. Chromosome arm 3L principal component analysis. (**A**) PC1 *vs.* PC2. (**B**) PC1 *vs.* PC3. Letter codes represent four sample collection locations: GHA = Ghana, GNB-A = Guinea Bissau – Antula (coastal), GNB-L = Guinea Bissau – Leibala (inland), GNB-S = Guinea Bissau – Safim (coastal).

Pairwise F_{ST} among all samples

Pairwise mean F_{ST} was recalculated, including the samples from Ghana. A clear dissonance can be seen, with low pairwise F_{ST} between coastal Guinea Bissau (Safim *vs.* Antula) and inland Guinea Bissau *vs.* Ghana (the lowest F_{ST} despite large geographical separation, ~1600km), yet higher F_{ST} in all coastal *vs.* inland comparisons (Figure 3.10). To allow ease of comparison, just 3L, giving a simplified but representative view of the autosomes, and X

chromosome arm means are shown in the figure. Figure 3.11 shows all chromosome arm F_{ST} means for comparison and highlights the strong effects of the 2La inversion polymorphism on the 2L chromosome arm. The highest mean pairwise F_{ST} is found in this region of the genome, between populations with high inversion orientation differences (Figure 3.11 and see also Figure 3.1).

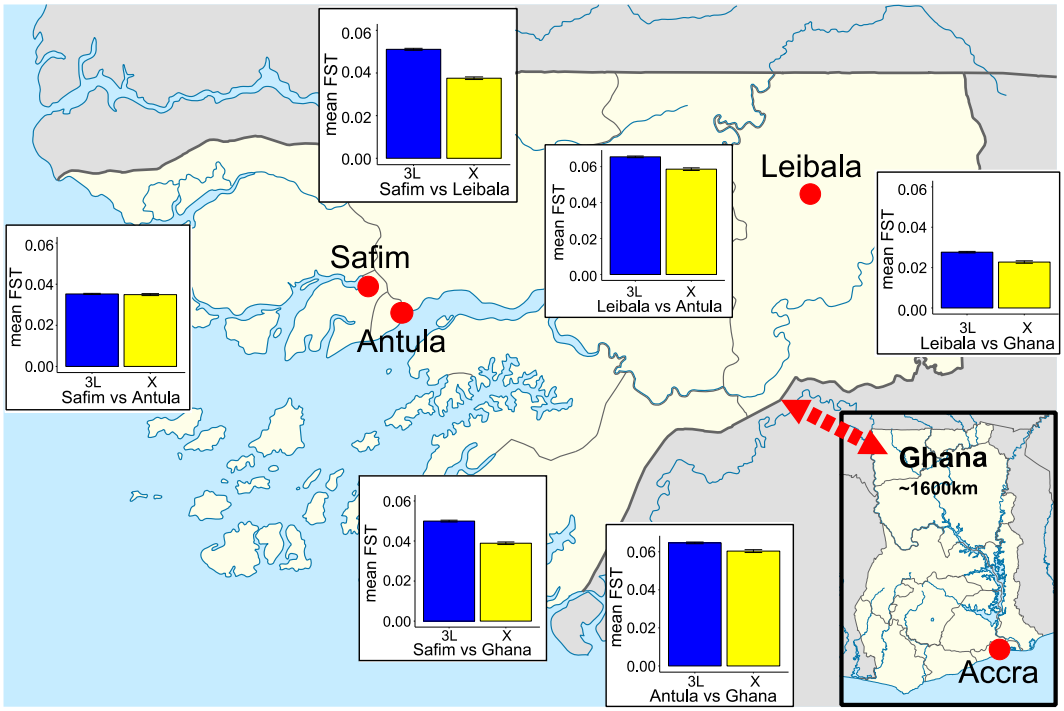


Figure 3.10. Map of Guinea Bissau showing mean pairwise F_{ST} between sampling locations. Locations of sample collection represented with red circles, inset map shows Ghana. Bar plots show mean chromosome arm F_{ST} for 3L and X with bars showing 95% confidence intervals.

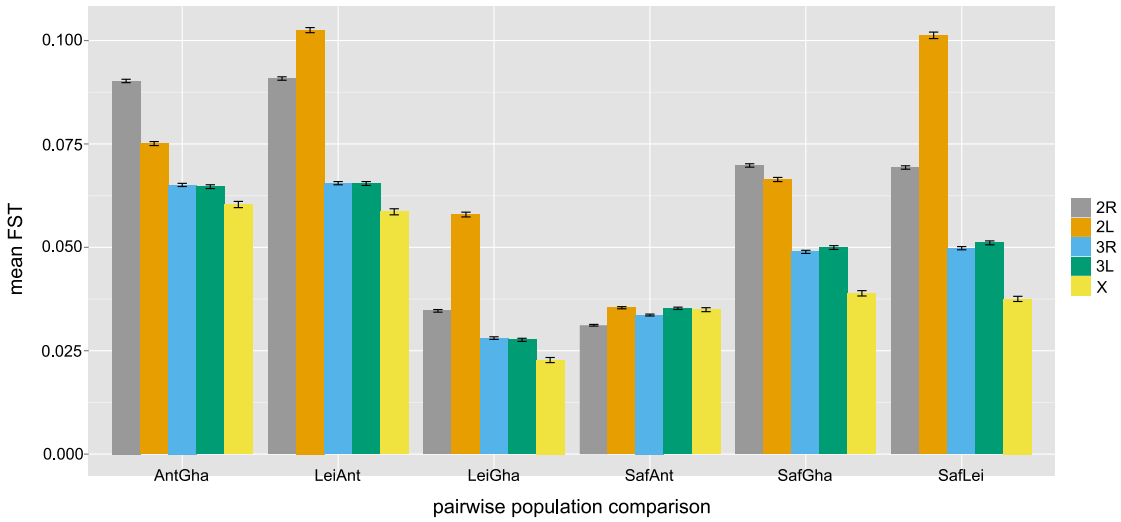


Figure 3.11. Chromosome arm mean pairwise F_{ST}s. Mean F_{ST} for all pairwise comparisons between Guinea Bissau and Ghana *A. gambiae* samples.

3.5 Discussion

3.5.1 Microsatellite analysis

Molecular analyses revealed a hybridisation hotspot in the coastal region of Guinea Bissau, with ~4 times more hybrids than found in the rest of the country (Table i1), or indeed found across the majority of the *A. gambiae*/*A. coluzzii* sympatric range (Vincente *et al*., unpublished;della Torre, Tu and Petrarca, 2005). These data reinforce previous findings of higher gene flow in the west of the range (Caputo *et al*., 2008; Oliveira *et al*., 2008), but show gene flow is not uniform across the country and highlight the coastal region as being the aberration. The addition of whole genome data for coastal and inland region reveals a hitherto unknown additional complexity to gene flow between these species.

3.5.2 Genomic analyses

Pairwise comparisons of inland and coastal Guinea Bissau *A. gambiae* reveal high levels of intra-specific divergence. Using a whole genome tiling array, Lee *et al.* detected divergence on the X chromosome between different chromosomal forms of *A. gambiae* in Mali (2013b) and suggested that adaptive X-linked genes might be driving the evolution and potential speciation of both chromosomal and molecular forms of mosquitoes. Though we find similarly high levels of intra-specific divergence, this is not restricted to the X chromosome but rather is genome-wide. The intraspecific divergence profile observed could be generated

by isolation, either via distance or the ecological zonation possibly underlying the three microsatellite clusters, but the topography of divergence resembles inter-species (*A. gambiae vs. A. coluzzii*) comparisons (Neafsey *et al.*, 2010), suggesting gene flow between the species as the cause.

3.5.3 Ancestry informative markers

In lieu of *A. coluzzii* WGS from Guinea Bissau, genome wide AIM analysis allowed elucidation of the potential inter-species introgression suggested in pairwise scans of differentiation. Results were striking: *A. gambiae* samples from the high gene flow coastal region showed a majority of markers affected by introgression from *A. coluzzii*, compared with inland *A. gambiae* samples, which showed less than 10% of markers introgressed. The AIMs corroborate the apparent *A. gambiae vs. A. coluzzii* topography of divergence seen in the F_{ST} scans. In the coastal region, asymmetric introgression from *A. coluzzii* had caused a genomic conversion across much of the *A. gambiae* genome.

The region containing the species-delineating markers on the X chromosome appears resistant to introgression. This may simply be due to its proximity to the X centromere (Noor and Bennett, 2009; Turner and Hahn, 2010), however a number of factors suggest that adaptive loci may present in this region and that a cost to fitness is incurred if this region introgresses. Firstly, this region of the X chromosome has recently been shown to be important in assortative mating (Aboagye-Antwi *et al*., 2015). Secondly, according to the species marker genotyping, the numbers of *A. coluzzii* in the coastal region appear to have declined. Sampling conducted in the coastal Antula saw a downward trend in *A. coluzzii* in collections between 1993 and 2010 (Gordicho *et al*., 2014). Approximately 25% of mosquitoes collected in 1993 were *A. coluzzii* but only 2.2% found there at the time of sampling took place for this study in 2010 (Vincente *et al.*, unpublished; Gordicho *et al.*, 2014). Therefore, the number of pericentromeric regions of *A. coluzzii* X chromosome (in which the species markers are found) has dropped in the region. However, high levels of *A. coluzzii* markers were found across the rest of the *A. gambiae* genome in the region (Vincente *et al.*, unpublished; Gordicho *et al.*, 2014). It appears that unlike the rest of the *A. coluzzii* genome, which has freely introgressed into *A. gambiae* and may therefore be adaptive or neutral, the last 3-4Mb of the X chromosome could be maladaptive and thus cannot successfully introgress.

It has recently been suggested that another large region of the anopheline X chromosome is also important in speciation. A genus scale study found a large proportion of the X chromosome (corresponding to the Xag inversion) to be highly resistant to introgression between *Anopheles* species (Fontaine *et al*., 2015). However, the inversion is monomorphic in *A. gambiae* and *A. coluzzii* so may not be relevant in speciation in this instance and unfortunately there are no AIMs in this region, so it is not clear how this part of the genome is reacting to high gene flow in Guinea Bissau. What is clear, however, is that in these high gene flow regions where species barriers have collapsed, single locus species markers are no longer useful, for example the species ratio in the coastal region may not be reliable with many more backcrossed hybrids than was detected. Though the X chromosome may be disproportionally important in speciation (Fontaine *et al*., 2015), a whole genome wide view is necessary here to characterise admixture in these populations and to define what species (or parts of species' genomes) are present in coastal Guinea Bissau.

3.5.4 Ghana calibration

Addition of *A. gambiae* samples from the typically low hybridization area of Ghana >1500 km away (della Torre, Tu and Petrarca, 2005), calibrated the topography on a wider geographic scale. Guinea Bissau inland samples are, in essence, genomically characteristic of pan-African *A. gambiae* (very low Ghana *vs.* inland divergence), whereas those from the coastal region strongly resemble *A. coluzzii* (as revealed by AIMs) However, without more widespread sampling and sequencing, it is difficult to ascertain where, geographically, the gene flow into the coastal *A. gambiae* is coming from. Very few *A. coluzzii* were recorded in the coastal region, though it is now clear that those recorded may actually also be hybrids. Gene flow may be high and completely contained within the coastal region, in which case the number of *A. coluzzii* (or rather their peri-centromeric X regions) may continue to decline. Alternatively genetic material may flow, uni-directionally towards the coast from the predominantly *A. coluzzii* central Guinea Bissau region, with continued reintroduction of *A. coluzzii* genetic material. In either scenario, these data suggest gene flow is asymmetric with genomic material introgressing from *A. coluzzii* into *A. gambiae.* An important question to now ask is what drives this high gene flow. The species have important phenotypic differences (Lehmann *et al.*, 2008); is this introgression pattern in the coastal region adaptive, e.g. in response to a suboptimal environment, or are the high levels of introgression we observed a side effect of another process such as a lack of reinforcement during key stages of divergence (Servidio and Noor, 2003)?

3.5.5 Future work

Though these data reveal comprehensively the dynamic gene flow environment in Guinea Bissau, there are still many unanswered questions. The sequencing of coastal, central and non-Guinea Bissau *A. coluzzii* would allow comparative genomics with the *A. gambiae* data to determine, spatially and genomically, where the asymmetric introgression originates and to discover if *A. coluzzii* genomes in the high hybridisation region still resemble that characteristic of the species. Collection on a micro-geographical scale in the coastal region may also help determine if introgression is adaptive. Strikingly similar within-species inland/coastal differentiation has also been noted in microsatellite data from The Gambia, another high gene flow region (Caputo *et al*., 2014). Expanding the WGS study to augment Gambian microsatellite data could expose potential general effects of coastal/inland partitioning and gene flow. The transect sampling regime employed in the collection of species and microsatellite data also opens the possibility of layering environmental data to investigate how extrinsic factors affect species composition and gene flow using landscape genetics techniques (Lowry, 2010).

3.5.6 Conclusions

In Guinea Bissau, gene flow rather than isolation appears to underpin observed differentiation, with the collapse of *A. gambiae* and *A. coluzzii* species rather than their speciation, as seen elsewhere in Africa, a more obvious outcome. The resultant coastal *A. gambiae* populations, although still presenting at least to some extent, the species' diagnostic markers, ostensibly appear to be *A. coluzzii* at a genome-wide scale. Further sampling is required to determine if this breakdown of species barriers is transitory or if homoploid hybrid speciation through genome sharing is occurring. The latter may necessitate re-defining of the major vector species in Guinea Bissau.

3.7 Acknowledgements

Tiago Antão (Liverpool School of Tropical Medicine) provided the SNP filtering parameters based on his wealth of experience filtering an *Anopheles gambiae* crosses experiment and the Ag1000g project (these have not yet been published). The figures and data relating to Guinea Bissau microsatellites and species proportions in the introduction are unpublished and come from collaborators on the project that this thesis chapter is based upon Joao Pinto (Instituto de Higiene e Medicina Tropical) and Bruno Gomes (Liverpool School of Tropical Medicine).

The differentiated SNPs, originating from the Neafsey *et al*., 2010 paper, were extracted by Giordano Botto (Wellcome Trust Centre for Human Genetics, Oxford).

3.8 Appendix

Appendix 3.8.1. Individual statistics

Appendix 3.8.1. Individual statistics. Mean sequencing depth calculated using autosomes.

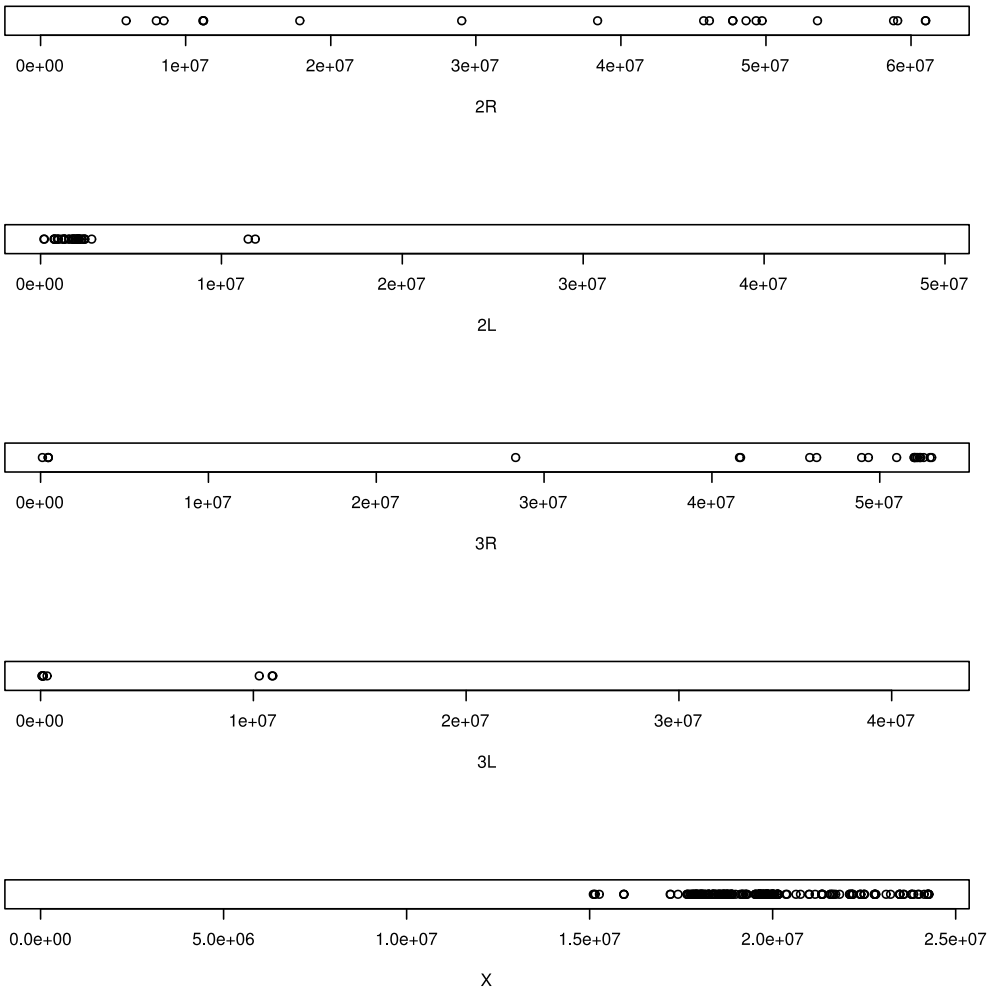
Individual code	Country	Region	Mean sequencing depth
AJ0001-C	Guinea Bissau	Leibala	29.8
AJ0007-C	Guinea Bissau	Leibala	26.6
AJ0009-C	Guinea Bissau	Leibala	31.5
AJ0011-C	Guinea Bissau	Leibala	29.8
AJ0013-C	Guinea Bissau	Safim	32.7
AJ0014-C	Guinea Bissau	Safim	20.8
AJ0016-C	Guinea Bissau	Safim	33.4
AJ0018-C	Guinea Bissau	Safim	22.7
AJ0020-C	Guinea Bissau	Safim	25.4
AA0006-C	Ghana	Greater Accra	24.7
AA0007-C	Ghana	Greater Accra	21.9
AA0008-C	Ghana	Greater Accra	24.9
AA0009-C	Ghana	Greater Accra	17.8
AJ0043-C	Guinea Bissau	Antula	27.9
AJ0047-C	Guinea Bissau	Antula	20.2
AJ0059-C	Guinea Bissau	Antula	31.9
AJ0061-C	Guinea Bissau	Antula	25.7
AJ0071-C	Guinea Bissau	Antula	34.1
AJ0076-C	Guinea Bissau	Antula	35.4
AJ0085-C	Guinea Bissau	Antula	87.2
AJ0096-C	Guinea Bissau	Antula	35.1
AJ0098-C	Guinea Bissau	Antula	33.6
AJ0100-C	Guinea Bissau	Antula	32.5
AJ0107-C	Guinea Bissau	Antula	29.7
AJ0113-C	Guinea Bissau	Antula	31.7

Appendix 3.8.2 Variants

Appendix 3.8.2. Number of variants across data sets. These values represent the number of sites remaining post quality control filtering and where sites with any missingness are removed. The Guinea Bissau column represents the "Guinea-Bissau-missingless" data set and all other columns the "All-sample-missingless" data set.

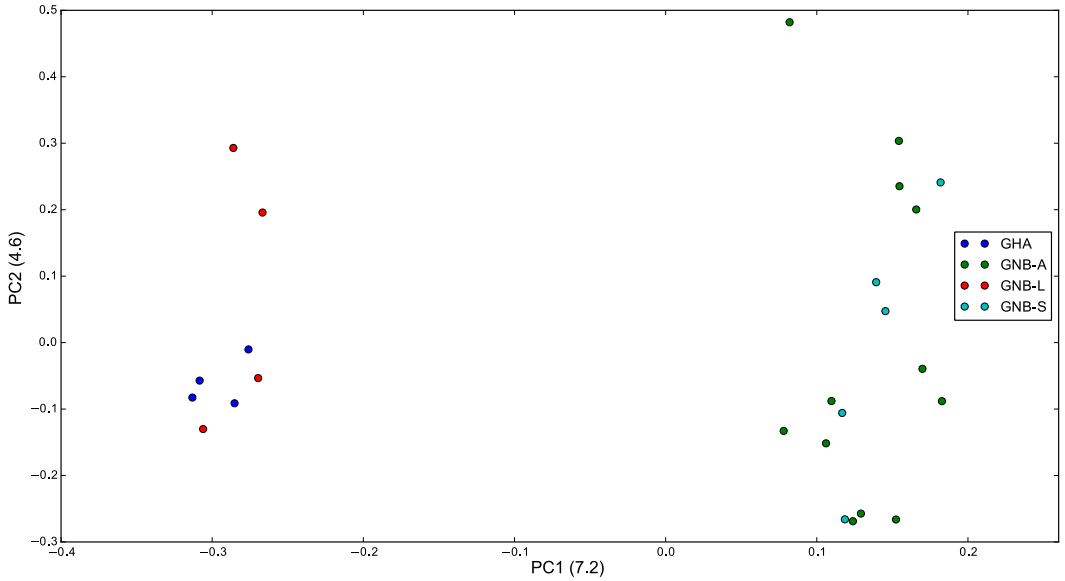
Chromosome arm	Just Guinea Bissau	All-sample individuals	All-sample singletons	All-sample no singletons
2R	1962624	727625	180790	546835
2L	1376521	464666	122614	342052
3R	1631746	563056	151787	411269
3L	1167925	386845	105060	281785
X	605874	138241	40631	97610
Whole genome	6744690	2280433	600882	1679551

Appendix 3.8.3 AIM positions



Appendix 3.8.3. Distribution of ancestry informative markers across chromosome arms. Horizontal bars represent chromosome arms and open circles reveal locations of ancestry informative markers.

Appendix 3.8.4 3R PCA



Appendix 3.8.4. Chromosome arm 3R principal component analysis. Letter codes represent four sample collection locations: GHA = Ghana, GNB-A = Guinea Bissau – Antula (coastal), GNB-L = Guinea Bissau – Leibala (inland), GNB-S = Guinea Bissau – Safim (coastal).

Chapter 4

Evolution of the *Anopheles gambiae* voltage gated sodium channel gene: a haplotype network approach

4.1 Abstract

With increasing resistance to insecticides, malaria vector control campaigns' successes are threatened. Resistance to the pyrethroid insecticides is of particular concern, as they are the only functional class approved for use on insecticide treated bednets. Pyrethroids, along with dichlorodiphenyltrichloroethane (DDT), target the voltage gated sodium channel (VGSC), an essential component of the mosquito nervous system. One evolutionary route for insecticide resistance is point mutations within the gene that disrupt the toxin binding to the active site. In *Anopheles gambiae*, three of these "*kdr*" mutations have been identified and explicitly associated with a resistant phenotype. Here we use whole genome sequencing of hundreds of individual *A. gambiae* genomes by the Ag1000G Consortium and a haplotype network approach to, for the first time, investigate whole gene genetic variation. Multiple origins of *kdr* mutations were evident but in a novel finding, long distance gene flow of resistant VGSC haplotypes was also shown to be an important factor in the evolution of insecticide resistance across Africa. Visualisation of these data in a network also revealed an abundance of high frequency non-synonymous mutations, unexpected due to the functional constraint suggested by conservation across dipterans. More striking was the non-random distribution of these protein altering mutations, most of which occurred on haplotypes carrying known resistance alleles and suggest additive or compensatory changes. Results show the complexity of the evolution of target site resistance, reveal a technique that can be used to identify, *in silico*, candidate resistance mutations and indicates that current molecular insecticide resistance assays may not elucidate the complexity of target site resistance in insects.

4.2 Introduction

Between 2000 and 2013 a 54% drop in deaths from malaria in sub-Saharan African (where ~90% of global cases occur) has been recorded, with the successful rollout of insecticide-based vector control campaigns being a key component (World Health Organization, 2011; World Health Organisation, 2014). However, the widespread use of insecticides creates a strong selective pressure for the evolution of insecticide resistance (Lynd *et al.*, 2010), a problem compounded by the limited range of insecticides available. As the only class approved for insecticide treated bednet (ITN) use, and also commonly used for indoor residual spraying (IRS), pyrethroids are the most medically important class of insecticides (van den Berg *et al*., 2012). Increasing resistance to pyrethroids could cause total failure of vector control campaigns (World Health Organization, 2012). With pyrethroid resistance becoming increasingly common (Ranson *et al*., 2011; Silva *et al.*, 2014), resistance mutations being shown to evolve *de novo* in the absence of gene flow (Reimer *et al*., 2005) and resistance already rendering ITNs ineffective in some regions (N'Guessan *et al*., 2007; Toé *et al.*, 2014), it is paramount that the evolution of insecticide resistance is understood for vector control campaigns to be targeted effectively to minimise further loss of efficacy (World Health Organisation, 2012; Jones *et al*., 2013).

4.2.1 Evolution of insecticide resistance

Both pyrethroids and DDT (the first widely used adult vector control pesticide) have a similar mode of action. These insecticides are neurotoxic, targeting the voltage gated sodium channel (VGSC), a protein present in the membranes of central and peripheral nervous system cells (Davies *et al*., 2007a). The insecticides bind to the channel causing it to remain open, stimulating hyper excitability and causing paralysis or "knockdown" before ultimately causing death (Davies *et al*., 2007a). One route to evolve resistance is by conformational changes to the protein structure of insecticide binding sites, via non-synonymous point mutations. First detected in the housefly, *Musca domestica*, these resistance associated variants in the VGSC DNA are known as knockdown resistance (*kdr*) mutations (Williamson *et al*., 1996). Multiple *kdr* variants have since been discovered across Insecta (Davies *et al*., 2007a), with two widespread in African *Anopheles* malaria vector mosquitoes (Ranson *et al*., 2011; Silva *et al*., 2014), *Vgsc-1014F* (Martinez-Torres *et* al., 1998) and *Vgsc-1014S* (Ranson *et al*., 2000). Another, more recently discovered *kdr* variant, *Vgsc-1575Y* is currently restricted to more westerly *Anopheles gambiae* and *A. coluzzii* populations (Jones *et al*., 2012a; Silva *et al*., 2014). The VGSC is a gene not only of great medical importance, but also a valuable evolutionary model. The great anthropogenic selection pressure due to

insecticide is both strong and recent (DDT use began in the 1940s, pyrethroids 1970s), allowing study of the emergence of adaptive variants in contemporary timescales (Davies *et al*., 2007a).

4.2.2 The voltage gated sodium channel

The physiological target of DDT and pyrethroids, the insect VGSC protein, is integral to the nervous system, allowing transmission of nerve impulses from neuron to neuron. Four domains make up the channel, each composed of six helical trans-membrane units including a voltage sensing unit and a unit which lines the interior of the ion channel as the four domains fit together on the cell (Davies *et al*., 2007a) (Figure 4.i1). The channel is controlled by two gates; during an action poential (nerve impulse) the change in voltage across the cell membrane (depolarisation) causes the protein channel to change shape, allowing Na+ (sodium ions) to pass from outside the cell down a concentration gradient to the lower Na+ concentrations found within , this is known as the 'm-gate'. The channel is then closed by the inactivation particle (the 'h-gate') which blocks the channel before the m-gate conformation is restored as the resting voltage difference across the membrane returns (Davies *et al*., 2007a).

The gene was discovered within the paralysis locus in *Drosophila melanogaster*, because of this, it is also referred to as 'para' (Loughney, Kreber and Ganetzky, 1989). In the sibling species *A. gambiae* and *A. coluzzii* , 31 exons code for the large gene (Giraldo-Calderón *et al*., 2014), which, situated centromere proximate on the 2L chromosome arm, lies within a large region of divergence detected between the species (Turner *et al*., 2005). The target of DDT and pyrethroid insecticides is the binding pocket, composed of IIS4-S5 linker and IIS5-III6 helices and two of the target site mutations found in these malaria mosquitoes (*Vgsc-1014F*/*S*) are situated on III6 helices, where they are thought to interact with the pocket reducing insecticide binding (O'Reilly *et al*., 2006). Interestingly, the only other *kdr* mutation that has been detected in *A. gambiae* is found in the III-IV linker, near the region coding for the inactivation particle, where its role conferring resistance is less clear (Jones *et al*., 2012a).

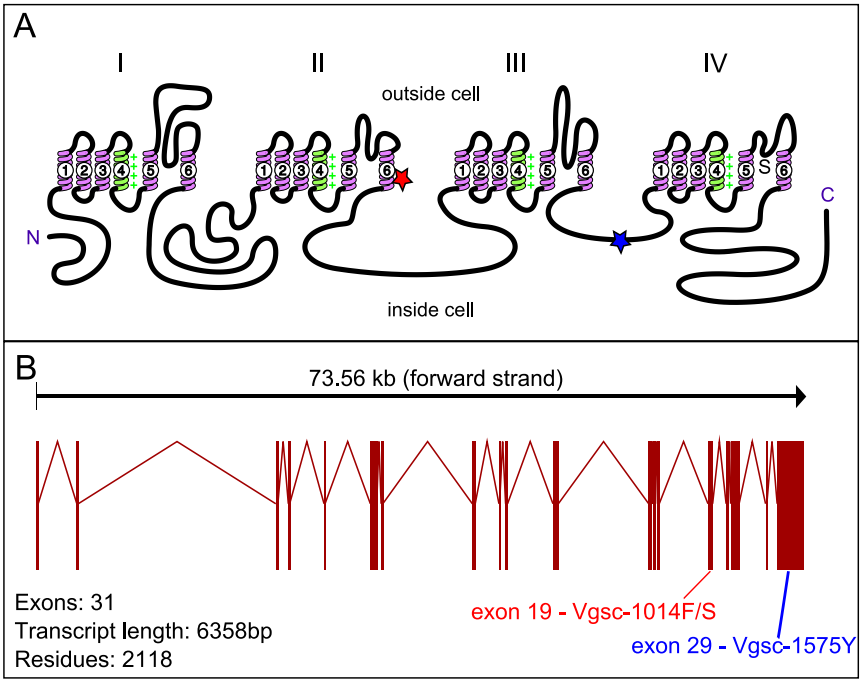


Figure 4.i1. The voltage gated sodium channel. (A) Cartoon represents the transmembrane structure of the protein, a single polypeptide chain. The four trans-membrane domains, I-IV, are composed of six (numbered) helices, which assembled make up the central channel pore (green helices form voltage sensing domains). Stars show approximate positions of known insecticide resistance target mutations (red – *Vgsc-1014F/S*, blue *Vgsc-1575Y*). Adapted from https://commons.wikimedia.org/wiki/File:Sodium-channel.svg – Creative Commons. (B) Exon/Intron structure of the VGSC gene with exons (vertical bars) containing known insecticide resitance mutations labelled. Adapted from VectorBase (Giraldo-Calderón *et al.*, 2014).

4.2.3 Gene genealogies: towards understanding the evolution, history and movement of insecticide resistance mutations

Investigating the evolution of interspecific genealogies has traditionally used bifurcating trees to explore relationships, however when comparing gene evolution within species or populations, known as tokogenies (Hennig, 1966), the assumptions of these approaches can be violated. Within taxa, ancestral haplotypes are often extant; representing these in a traditional tree is problematic (Templeton, Crandall and Sing, 1991), and multiple mutations from the ancestral haplotype can result in multiple descendent haplotypes. These multifurcations occurring alongside reticulations, which can be caused by recombination, cannot easily be represented by bifurcating trees (Posada and Crandall, 2001). Moreover, because intraspecific genealogies can be much less diverse than in comparisons to genes

across taxa, statistical support may be limited (Posada and Crandall, 2001). Fortunately using a network approach, rather than a tree, accommodates many such population-level idiosyncrasies. Gene or haplotype networks allow for phylogenetic relationships in intraspecific data to be easily visually explored. Coalescent theory may also be utilised to infer directionality and resolve ambiguous network connections (Templeton, Crandall and Sing, 1992; Crandall and Templeton, 1993)

4.2.4 Aims

Here we exploit the power of the *Anopheles gambiae* 1000 Genomes dataset to produce a high resolution network view of variation within a gene of great importance in the fight against malaria, the VGSC.

- The coding region of the gene is conserved across dipterans, suggesting high levels of functional constraint/purifying selection (Davies *et al*., 2007a), leading to a prediction of a relative imbalance of synonymous to (function altering) non-synonymous variants. This prediction will be explored.
- Evidence from other species of insects suggests *kdr* mutations are deleterious (Foster *et* al., 2003; Brito *et al*., 2013), yet they have been shown to sweep rapidly through populations (García *et al*., 2009; Lynd *et al*., 2010). Therefore, evolutionary forces to evolve, not just stronger resistance to insecticides but compensatory mutations to ameliorate the deleterious fitness effects of such changes to an essential nervous system protein may be being exerted upon insecticide pressured vector populations. The diversity of non-synonymous mutations in the VGSC and evidence for their importance will be investigated.
- Few origins of *kdr* mutations have been revealed in previous studies (Pinto *et al.*, 2007; Etang *et al*., 2009), yet it has been shown that these resistance variants can evolve *de novo* in isolation from gene flow (Reimer *et al*., 2005). Utilising the power contained within the large dataset in a network approach will enable the relative contributions of gene flow and new origins of the mutations to be evaluated on a whole gene pan-African scale for the first time.

4.3 Methods

4.3.1 Ag1000G

For detailed information on sample collection, sequencing, variant calling and quality control, see (Ag1000G, 2015). In brief, *A. gambiae* and *A. coluzzii* females were collected from eight countries across Africa:Angola, Burkina Faso, Cameroon, Gabon, Guinea, Guinea Bissau, Kenya, Uganda. Guinea Bissau was an *A. gambiae/coluzzii* admixed population and Burkina Faso had collections of both species. Samples were whole genome sequenced using Illumina technology. Sequenced reads were aligned to the *A. gambiae* AgamP3 reference genome assembly (Holt *et al.* 2002), then aligned bam files underwent improvement, before variants were called and filtered. AR2 filtering rules removed variants if: FS < 60.0, HRun > 4, DP <18000 or >32000, MQ < 40, alleles > 2, QD < 5, ReadPosRankSum < -8, overlapped positions repeat masked by DUST (http://blast.wustl.edu/pub/dust). This process produced a call set containing 765 individuals which was released only to the *Anopheles gambiae* Thousand Genomes (Ag1000G) Consortium. Before publically releasing the data, it was decided that more confidence in variant calls could be gained by the addition of another layer of filtering. This produced the more conservative AR3 data set which used the the same filters as AR2, plus variants were removed if: the number of samples with high coverage – more than twice mode for chromosome > 15 samples, low coverage – less than half mode for chromosome > 76 samples, missing data > 1 sample, reference base = N, number of samples with average MQ below 30 > 76, number of samples with more than 10% of reads with MQ0 > 1 sample). Both call sets were used here, AR2 being used to investigate an insecticide resistance variant that was present in that earlier release but was removed by the AR3 filtering rules.

4.3.2 Variant extraction

All voltage gated sodium channel (VGSC) gene variants in both the Ag1000G phase 1 AR2 and AR3 (public release) data set were extracted using VCFtools 0.1.12a (Danecek *et al.* 2011), with exon position information in .bed format downloaded from the VectorBase AgamP4 genome assembly (Megy *et al.* 2012)(gene ID AGAP004707). For the network analysis, to allow ease of visualisation by reducing the number of nodes, singleton variants were removed (vcftools –gzvcf <input.vcf.gz> --bed <exons.bed> –maf 0.001307 --recode --out <out.vcf>). Missingness within the data set required a higher minor allele (MAF) cut-off than the 1/1530 expected (750 diploid individuals = 1530 haplotypes) to remove all singletons (see "–maf" above). Biological signals in singleton data were investigated separately (see below).

4.3.3 Phase and network

To identify and investigate variation and track the origins of VGSC haplotypes, the genomic data must be phased. The combined homologous chromosome information (with homozygous and heterozygous positions) is separated into maternal and paternal contributions. Here, the Bayesian coalescent algorithm, Phase 2.1.1 (Stephens *et al.* 2001, Stephens and Scheet 2005), was used to infer haplotypes (./PHASE <input.inp> <output>). In order to change the .vcf variants into Phase input format, a custom Perl script was used (https://github.com/cclarkson/thesis_chapter_4/ blob/ master/VCF_to_PHASE_input.pl). The haplotypes were initially visualised as a network using the statistical parsimony algorithm implemented in TCS 1.21 (Clement et al. 2000) by re-formatting the Phase output haplotype and frequency data using a custom Perl script (https://github.com/cclarkson /thesis_chapter_4/blob/master/Phase_to_phy_input.pl). Due to high divergence between haplotypes, TCS 'Fix Connection Limit At' was set at 500 to allow for all node inclusions in a single network regardless of how diverged. It should be noted however, that because the highly divergent haplotypes are difficult to incorporate parsimoniously, their positioning exceeds the theoretical limits of the approach so may be not be correct. The network was exported from TCS in .gml format allowing visualisation and addition of pie charts in Cytoscape 3.1 (Smoot *et al.* 2010).

4.3.4 Extended haplotype homozygosity

To explore how the landscape of selection affects *kdr* and non-*kdr* bearing haplotypes differently (*Vgsc-1014F/S*), extended haplotype homozygosity (EHH) analysis was employed. EHH allows comparison of fine scale linkage disequilibrium (LD) decay either side of a defined 'core' haplotype region (Sabeti *et al*., 2002; Sabeti *et al*., 2006). In this case, as in a previous study by Lynd *et al*. (2010), LD decay was compared in both telomeric and centromeric direction either side of the *kdr* mutations' location in both wild type and 'resistant' haplotypes. EHH analysis requires phased data so the VGSC exonic haplotypes inferred using Phase (see above) (Stephens *et al.* 2001, Stephens and Scheet 2005) were split into three categories, either *Vgsc-1014F*, *Vgsc-1014S* or wild type *Vgsc-1014L* codon bearing. EHH was calculated and plotted separately for each of these categories using R code (R Development Core Team, 2014; Weetman *et al*., 2015). Significance each of EHH values was determined using 95% confidence intervals (CI) calculated by bootstrapping using 1000 replications, again using R (R Development Core Team, 2014; Weetman *et al*., 2015). Bifurcation plots were also produced, using the same data and the software package SWEEP

v1.1 (Sabeti *et al.*, 2002), to allow visualisation of the breakdown of LD either side of the core region; any SNP where two segregating alleles are found causes a bifurcation at that point.

4.3.5 Non-synonymous variants

The AR3 data set was annotated using SNPeff (Cingolani *et al.*, 2012) to enable the effects of individual variants *e.g.* coding/non-coding, synonymous/non-synonymous to be classified. Upon release of AR3, SNPeff annotations were used to verify the AR2 variants classed as non-synonymous using Ensembl's Variant Effect Predictor (VEP) algorithm (McLaren *et al.*, 2010); classification was fully concordant between methods.

4.3.6 Taqman

Non-synonymous variants annotated in the data set were compared to 36 VGSC variants previously associated with insecticide resistance across Insecta (Rinkevich *et al.*, 2013). Other than the three, well described *kdr* mutations previously discovered in *A. gambiae*/*coluzzii* (*Vgsc-1014F* (Martinez-Torres *et* al., 1998), *Vgsc1014S* (Ranson *et al.*, 2000), *Vgsc-1575Y* (Jones *et al*., 2012a)), two variants were found which altered the 1879 codon of the VGSC, a codon where variants had been linked to resistance in the diamondback moth *Plutella xylostella* (Sonoda *et al.*, 2008). High throughput TaqMan (Life Technologies) SNP detection assays were designed for these two polymorphisms at VGSC 1879 (P1879L, P1879S) to enable genotyping of resistance phenotyped individuals and resistance association tests to be conducted.

VCFtools (Danecek *et al.* 2011) was used to extract AR2 variants (the assay design was conducted pre-AR3 release) approximately 200bp 5' and 3' of the 1879 residue (2L:2430657–2431079). These variants were to inform the design of the primers and probes, to ensure that they would not be designed to bind known variable regions. A custom Perl script was used to reconstitute a consensus sequence with ambiguity codes from the .vcf (https://github.com/cclarkson/thesis_chapter_4/blob/master/ Variant_positions_from_vcf.pl). Ambiguity codes were transferred into the AgamP3 reference sequence for the region by aligning both in a text editor (Appendix 4.6.1). Because two assays were required, one for each of the 1879 mutation and due to the proximity of these two 1879 variants (consecutive bases), it was not possible to annotate the 'other' 1879 mutation in each design therefore

when submitting the sequence to Life Technologies for assay design, in the P1879L assay the reference allele had to be in place in the P1879S position (and *vice versa*) rather than the ambiguity code, N, which should have been in place for variant site; the probe cannot be designed to bind a variant region.

The 1879 TaqMan assays were run on DNA from females from the Tiassalé colony strain of *A. gambiae* which had undergone WHO bioassay (World Health Organisation, 2013) with deltamethrin (0.05%); two separate sets of tests were conducted, the first from assays with alive *vs.* unexposed females and a second with alive *vs.* dead females. This strain was colonised from Côte d'Ivoire in 2013 and underwent insecticidal selection pressure every six months with deltamethrin, a pyrethroid insecticide (Bagi *et al.*, 2015). DNA was extracted using from individuals using DNeasy Blood and Tissue Kit (Qiagen). A Strategene MX3005P thermal cycler (Agilent Technologies) was used with reaction [TaqMan Gene Expression Master Mix (Applied Biosystems) : 5µl, probe (see above) : 0.125µl, DNase free water : 3.875µl, DNA : 1µl] and a thermal cycle [92°C : 10min (92°C : 15s, 60°C : 60s) x 40 cycles]. All samples were also genotyped for the *Vgsc-1014F kdr* mutation known to be present in the Tiassalé colony using the TaqMan assay described in Bass *et al.* (2007). The *Vgsc-1014S* and *Vgsc-1575Y* (Jones *et al.*, 2012a) *kdr* mutations are absent from this colony

4.3.7 Sanger sequencing

To quality assure the TaqMan assay, Sanger sequencing was conducted both on PCR product and on cloned DNA. The genomic region surrounding the 1879 locus was amplified using forward primer: AGGGCTATCCGGGAAATTGT, reverse primer: GTACTCTTCACGCTG CCTCC, with reaction concentrations [Dream Taq (Thermo Scientific) : 0.2µl, buffer : 2.5µl, dNTPs : 0.5µl, forward primer : 0.5µl, reverse primer : 0.5µl, DNase free water : 19.8µl, DNA : 1µl] and a thermal cycle of [94°C : 5min, (94°C : 30s, 60°C : 30s, 72°C : 60s) x 30 cycles, 72°C : 5min]. A 2720 Thermal Cycler from Applied Biosystems was used and nine females from the Tiassalé colony were sequenced. Cloning (CloneJET PCT Cloning Kit - Thermo Scientific) and Sanger sequencing was also used to validate 'double heterozygote' female individuals with ambiguous genotypes (to determine if both 1879 mutations were found on the same haplotype). Three colonies from each of 12 female Tiassalé individuals displaying the ambiguous genotype were sequenced using the same primers as before. Traces were examined, cleaned and genotypes, for the 54 successfully sequenced haplotypes, were determined using the Codon Code Aligner software (http://www.codoncode.com).

4.3.8 Haplotypic insecticide resistance association tests

To compare associations between phenotypic outcomes from insecticide resistance bioassays (alive/dead/control) (World Health Organization, 2013) and the haplotypes generated from the two codons (*Vgsc-1014* and *1879*) determined byTaqMan assays, odds ratios were calculated. Haploview v4.1 was used to produce haplotype associations: *Vgsc-1014F* and *Vgsc-1879S, Vgsc-1014L* and *Vgsc-1879S*, *Vgsc-1014F* and *Vgsc-1879L, Vgsc-1014L* and *Vgsc-1879L* (Barrett *et al*., 2005), from which the odds ratios for survival in the presence of insecticide were determined.

4.3.9 Phylogenetic tree

Investigation into possible introgression of VGSC alleles required us to use outgroup comparators. Representative haplotypes, both 'susceptible' (*Vgsc-1014L*) and 'resistant (those bearing *Vgsc-1014F* and *S*), were selected from nodes across the network (Appendix 4.6.2). Due to high levels of conservation of exons across *Anopheles* species, non-coding (intronic) sequence was included in the analysis to improve resolution. Intron 18 and exon 19 (the exon containing *Vgsc-1014* locus) variants were extracted from the network haplotypes and consensus nucleotide sequence was re-constructed for each using a custom perl script (https://github.com/cclarkson/thesis_chapter_4/blob/master/Create_consensus_from_PHASE.pl) and the *A. gambiae* reference sequence (AgamP4). Sequence from potential introgression partners *A. coluzzii* (AcolM1), *A. arabiensis* (AraD1), *A.melas* (AmelC2) and more distantly related *A. funestus* (AfunF1) and *A. christyi* (AchrA1), were extracted from Vectorbase by BLASTing the AgamP4 sequence and taking the top hit (Giraldo-Calderón *et al*., 2014). Alignment of these sequences using Mega v6 (Tamura *et al*., 2013), revealed very poor alignment of *A. funestus* and *christyi* due to divergence outside of the exonic region so these were excluded from the analysis. Good alignment, however, was found for 2000bp (including exon 19 and the preceding intronic sequence) for *A. melas*, *coluzzii*, and *arabiensis*. A maximum likelihood tree was constructed using these sequences, the Ag1000G haplotypes and the AgamP4 reference sequence, using Mega v6; 1000 bootstrap iterations were performed to determine the reliability of the topology (Tamura *et al*., 2013).

4.4 Results and Discussion

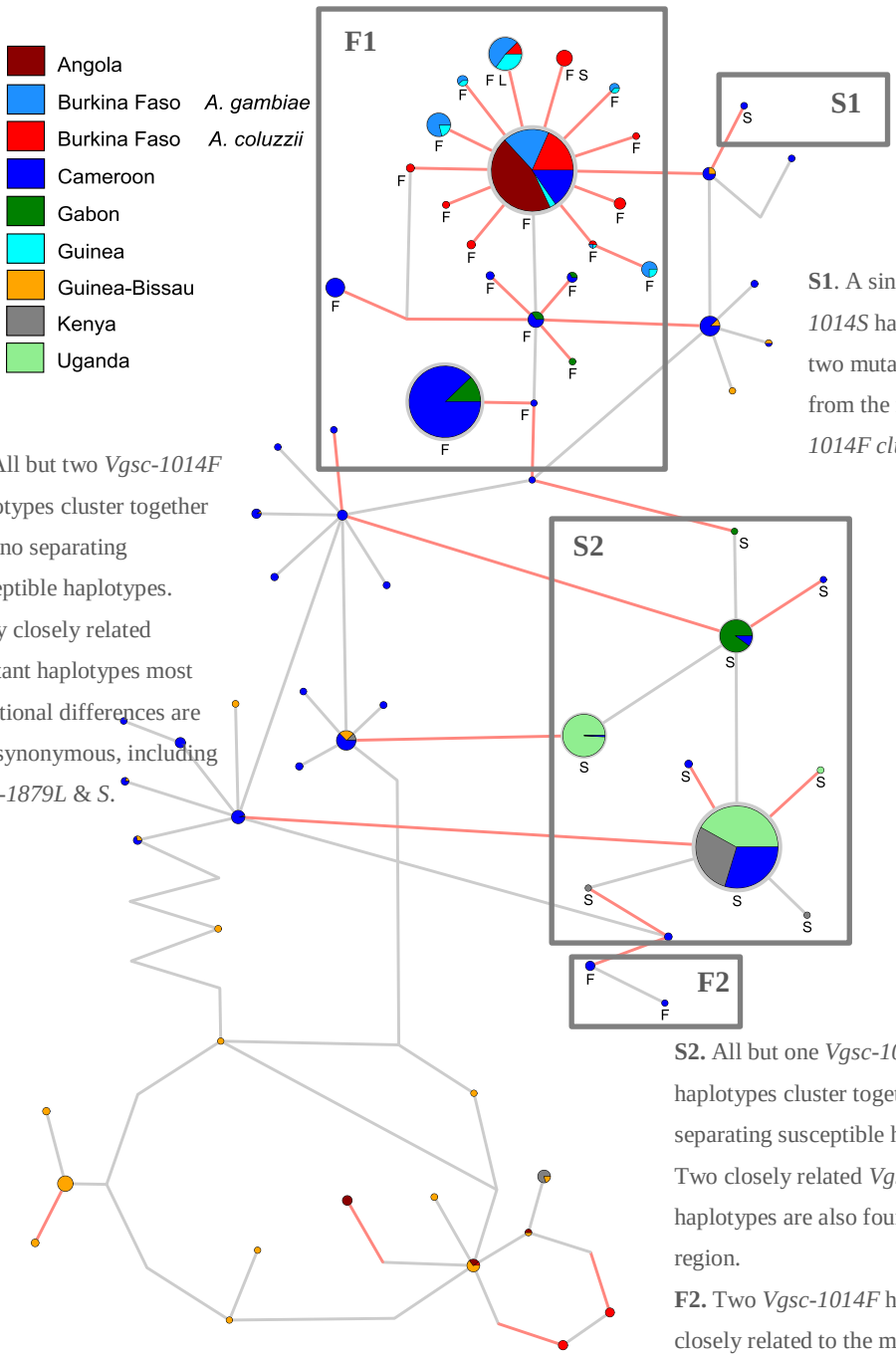
Visualisation of relationships within a large and complex genomic variation dataset like the pan-African VGSC gene analysed here is difficult. However, we have demonstrated that a haplotype network approach allows concise representation and understanding of this type of data, even with over 100 different haplotypes. Describing data in this way allowed for layering of meta data such as haplotype frequency, locations and nature of differences between haplotypes (synonymous or non-synonymous). These additional data enabled a new insight into the evolutionary history of a medically important gene, revealing that widespread *kdr* resistance is caused both by multiple origins and the long range gene flow of resistance mutation bearing haplotypes. The network approach also revealed unexpected high levels of non-synonymous diversity on resistant haplotypes, including some mutations linked to resistance in other insects, suggesting a layering of mutations may provide additive resistance (as found previously with the *Vgsc-1575Y* – Jones *et al*., 2012a) and/or be compensatory for the deleterious fitness effects of carrying *kdr* mutations (Foster *et* al., 2003; Brito *et al*., 2013). The ability to track and predict effects and movement of these mutations is essential for the planning and success of insecticide vector control campaigns and as these campaigns turn to large genomic data sets, haplotype networks can be employed to generate clear and intuitive visualisations.

4.4.1 Exonic network – non-synonymous mutations

Previous research into the *A. gambiae/coluzzii* VGSC had concentrated on variation contained in intron 19, non-coding data (Pinto *et al*., 2007; Etang *et al*., 2009; Santolamazza *et al*., 2015). Variation in nucleotide sequence was described, and in one case high levels of variation found (Santolamazza *et al*., 2015). However, this variation was perhaps unsurprising given the reduced purifying selection on non-coding regions, furthermore, these studies of variation could not reveal functional differences between haplotypes. Here we identify a surprising amount of previously undetected non-synonymous mutations across the haplotypes, some at high frequencies and spanning a large geographical range.

Analysis of coding variation in the VGSC revealed an abundance of seemingly non-randomly distributed non-synonymous variants. Using the edges (inter-node connections that show single nucleotide differences) between the nodes of the network (Figure 4.1), ratios of non-synonymous (N) to synonymous (S) mutations were estimated for different resistance "classes" of haplotypes and could be compared. For variants found on a *Vgsc-1014F*

possessing ‘resistant’ haplotype N/S = 26/4, for those on a *Vgsc-1014S* ‘resistant’ haplotype N/S = 9/5 and for those on a 1014L ‘susceptible’ haplotype N/S = 5/52; a contingency test showed that these values are significantly different from each other (x^2 = 54, df = 2, p = <0.0001). This protein altering variation was surprising both because of the presumable functional constraint on the VGSC gene (Davies *et al*., 2007a) and because of the distribution of the mutations being commonly found on haplotypes already bearing a known *kdr* mutation, particularly *VGSC-1014F*.



F1. All but two *Vgsc-1014F* haplotypes cluster together with no separating susceptible haplotypes. Many closely related resistant haplotypes most mutational differences are non-synonymous, including *Vgsc-1879L & S*.

S1. A single *Vgsc-1014S* haplotype just two mutational steps from the main *Vgsc-1014F cluster*.

S2. All but one *Vgsc-1014S* haplotypes cluster together with no separating susceptible haplotypes. Two closely related *Vgsc-1014F* haplotypes are also found in this region.

F2. Two *Vgsc-1014F* haplotypes closely related to the main *Vgsc-1014S cluster*.

Figure 4.1. Haplotype parsimony network of VGSC exonic variation. Statistical parsimony network composed of phased AR3 haplotypes. Node size represents relative frequency of haplotypes, F = L1014F carrying haplotype, S = L1014S, F S = L1014F + P1879S, F L = L1014F + P1879L. Red edges denote non-synonymous nucleotide changes, grey edges synonymous. Pie charts show proportion of haplotypes from country/species which make up each node. Grey boxes denote clusters of *kdr* mutation containing haplotypes unseparated by 'susceptible' haplotypes, potential resistance mutation origins.

4.4.2 Resistance mutations

High numbers of non-synonymous mutations on known resistance backgrounds could be conferring an increased resistance to insecticides (Figure 4.1). These protein altering variants were screened to discover if any had been previously linked to resistance in other animals. Using a recent review of known arthropod VGSC mutations linked to resistance (Rinkevich *et al*., 2013), 36 codon positions across the gene were converted from *Musca domestica* codon numbering to *Anopheles gambiae* (Appendix 4.6.3) by aligning the *A. gambiae* VGSC protein sequence obtained from Vectorbase (AGAP004707, Giraldo-Calderón *et al*., 2014) to that of the *M. domestica* VGSC sequence (genbank: X96668) in Codoncode Aligner (http://www.codoncode.com). However, only two of the 25 non-synonymous VGSC mutations in the AR3 data lay within codons previously linked with insecticide resistance and these two variants caused different amino acid changes to the same codon, 1879 (in *A. gambiae* VGSC transcript RA this equates to position 1874, in transcripts RB and RC position 1848).

4.4.3 *Vgsc-1879* and insecticide resistance

The two variants previously linked with insecticide resistance, *Vgsc-1879L* and *Vgsc-1879S* were only found on haplotypes bearing *Vgsc-1014F* in West Africa. *Vgsc-1879L* was found in *A. gambiae* from Guinea and Burkina Faso and *A. coluzzii* from Burkina Faso. *Vgsc-1879S* was found only in *A. coluzzii* from Burkina Faso (Figure 4.1; Figure 4.2). The high frequencies of these haplotypes (n_L = 80, n_S = 29) and the fact that *Vgsc-1879L* is found across two countries and two species hint that these mutations may be under positive selection. These, and the other non-synonymous mutations found here may contribute to insecticide resistance and therefore be under positive selection to increase fitness in the presence of insecticide as seen for *Vgsc-1014F* (Lynd *et al*., 2010). The 1879 codon was previously linked to insecticide resistance in the crop pest diamondback moth *Plutella xylostella* in an association study (Sonoda *et al*., 2008).

To model structural chances produced by the insecticide resistance candidate 1879 mutations and therefore predict potential functional effects, we attempted to add the altered residue to an existing 3D model of the *A. gambiae* VGSC gene. However, codon 1879 falls in the final exon of the gene (exon 31), at the extreme C-terminal cytoplasmic tail, outside of the existing model and some distance from the insecticide binding site (E. Davies pers. comm.; Davies *et al*., 2007b).The remaining non-synonymous mutations that fell within the

modelled transmembrane regions were annotated using the model, however none fell in regions which are known to interact with active insecticide binding sites (E. Davies pers. comm.). Clearly further research is required to functionally characterise these potentially important variants as regions that may change insecticide effects would not necessarily need to physically interact with a binding pocket as shown by the *Vgsc-1575Y* mutation (Jones *et al*., 2012a).

Despite falling outside of the transmembrane region, high frequency tip nodes, haplotype presence across multiple countries (Figure 4.1) and association with resistance in other species (Sonada *et al*., 2008) make 1879 codon changing mutations resistance candidates. TaqMan assays (Life Technologies) were therefore developed to enable genotyping of insecticide phenotyped *A. gambiae.* To test for a link between presence of either *Vgsc-1879L* or *Vgsc-1879S* and pyrethroid resistance, mosquitoes from the Tiassalé colony (Cote D'Ivoire) were employed as all three 1879 alleles detected in Ag1000g data (L, S and P) were found to be segregating in the strain (TaqMan quality control – Appendix 4.6.4). Female mosquitoes underwent WHO bioassay with the pyrethroid insecticide deltamethrin (World Health Organisation, 2013) and mortality quantified 24 hours after exposure. Alive, dead and control (unexposed to insecticide) mosquitoes where then genotyped. A pyrethroid was used in the bioassay following the association test that demonstrated the link between the 1879 locus to resistance in *P. xylostella* using this class of insecticide (Sonoda *et al*., 2008).

Haplotypic association tests were used to explore relationships between bioassay phenotypes and genotypes determined by *Vgsc-1014L/F*, *Vgsc-1879P/L* and *Vgsc-1879P/S* TaqMan assays. Unfortunately, across two separate tests results were contradictory and inconclusive. In the first test 88 Tiassalé females, a mixture of bioassay survivors and unexposed individuals were genotyped. Significant additive effects of both *Vgsc-1879L* and *S* above the survivorship seen for *Vgsc1014F* were found suggesting that additional resistance is conferred by the presence of the 1879 mutations (Figure 4.2a). The odds ratios here are conservative due to bioassay survivors being compared to unexposed (control) individuals rather than bioassay dead individuals. However, results from a second test using 154 individuals from the colony (several generations after the first test) found no significant effects of the 1879 codon mutations, despite survivors being compared to dead individuals (did not survive a 1 hour insecticide exposure) (Figure 4.2b). Strong significance of the first test suggests resistance association but the assays clearly require repeating to confirm. There

may have been problems with the Tiassalé colony in the second test due to bottlenecks as the *Vgsc-1014F* had become fixed (why the second assay has one less haplotype – Figure 4.2b).

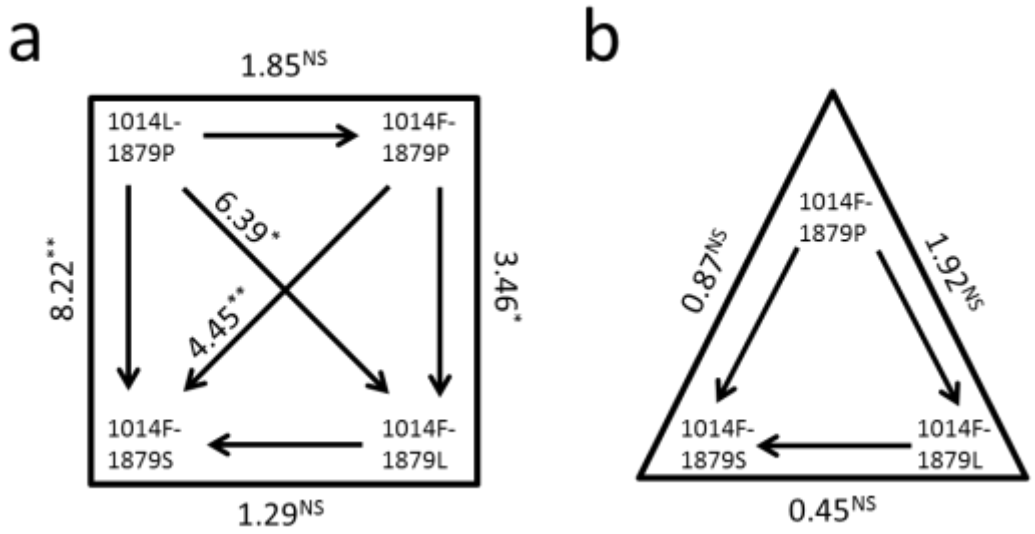


Figure 4.2. Summary of haplotypic association tests with for the allele combinations found at Vgsc-1014 and Vgsc-1879 with resistance phenotype to deltamethrin. Arrows denote direction of odds ration test with values showing the odds ratio and superscript denoting significance as tested with chi-square with Yates continuity correction (NS = not significant, * = ≤0.05, ** = ≤0.01). (**a**) Association test on 88 Tiassalé colony females, 47 unexposed to bioassay, 41 survivors of 1 hour exposure to 0.05% deltamethrin. (**b**) Association test on 154 Tiassalé colony females, 43 dead and 111 alive after 1 hour bioassay.

Another, not mutually exclusive, explanation for the high levels of non-synonymous mutations, and for their biased presence on resistant haplotypes, is the known deleterious nature of the *kdr* mutations (*Musca domesticus* - Foster *et* al., 2003; *Aedes aegypti* - Brito *et al*., 2013). In the presence of insecticides the *kdr* mutations confer a clear fitness advantage and increase in frequency but in absence of these toxins the resistance alleles quickly reduce in frequency (Brito *et al*., 2013). The abundance of non-synonymous variants on *kdr* containing haplotypes may be acting as compensatory mutations, ameliorating the deleterious effects of resistance on mosquitoes in regions pressured by insecticides. Compensatory mutations may also be enabling the trans-continental gene flow that is found in the resistant haplotypes, with some spanning many thousands of kilometres. With large differences seen in ITN and LLIN coverage on a continent wide scale (Flaxman *et al*., 2010;

Griffin *et al*., 2010), the heterogeneity of insecticide pressure at the local scale is presumably great. Yet, these resistant haplotypes appear able to cross vast geographical distances, potentially spending generations in mosquitoes un-pressured by insecticides, without resistant alleles being lost. The non-synonymous mutations on *kdr* haplotypes may be conferring compensatory effects for the fitness costs of *kdr* in insecticide absence (Foster *et* al., 2003; Brito *et al*., 2013), perhaps allowing these haplotypes to spread and increase in frequency. This generates a prediction of eventual replacement of 'pure' *Vgsc-1014F* haplotypes with the haplotypes carrying these additional non-synonymous mutations, leading to more resistant mosquitoes (compensatory mutations increasing overall fitness). However, it is important to note that the most widespread resistant haplotype (the largest*Vgsc-1014F* bearing node – Figure 4.1), does not have these 'extra' non-synonymous mutations suggesting that they may have occurred post-long-distance gene flow and therefore could be involved in the local adaptation of haplotypes.

Key to the hypotheses above is that with large population sizes (Pinto *et al*., 2003; Crawford and Lazzaro, 2010; Karasov *et al*., 2010), even large sample sizes of tens or hundreds of individuals, as found in this huge data set, must generally only detect haplotypes at high frequency in these populations. For example the striking absence of synonymous mutations in the star shaped *Vgsc-1014F* region of the network may be explained by the short timescales over which evolution is acting (<60 years - Davies *et al*., 2007a) in concert with the probability that only haplotypes carrying fitness enhancing non-synonymous variants (under positive selection from either hypothesis above or another driver) would increase in frequency quickly enough to stand a high chance of being detected by the relatively shallow sampling. With whole genome sequencing (WGS) providing the ability to identify variation across the entire VGSC gene and from many individuals, a surprising amount of non-synonymous mutations were found within this gene, presumably under high functional constraint. The network approach suggests that many of these variants may be driving haplotypic selective sweeps. Results show a previously unknown and unexpected level of VGSC complexity belied by the simple labelling individuals as either carriers of the three known *A. gambiae kdr* mutations or wild type. Understanding and tracking these "secondary" mutations, whether they be compensatory or conferring additional resistance, should now be a priority for vector control.

4.4.4 *Vgsc-1575Y*

Of the three previously described resistance conferring VGSC mutations, only two were found in the AR3 data set, *Vgsc-1014L* and *Vgsc-1014S* (Martinez-Torres *et al*., 1998;

Ranson *et al*., 2000). A third, *Vgsc-1575Y*, was conspicuous by its absence (Jones *et al*., 2012a). Though possible that no samples with the variant were sequenced, the mutation has previously been documented as widespread in West and Central Africa including two of the countries sampled in Ag1000G, Burkina Faso and Cameroon (Jones *et al*., 2012a). To investigate whether the absence of this mutation was a true biological signal or one resulting from stringent data filtering, the raw unfiltered AR3 data set was examined. The variant position causing the amino change determined from protein alignment with *Musca domesticus* was 2L:2429745 (not the position incorrectly reported in Jones *et al*., 2012a), and this was present as a bi-allelic variant in the raw AR3. It appears that although the insecticide resistance associated allele frequency was 104, it just failed the "HighCoverage" quality control filter for inclusion. The filter is based upon coverage depth at a genomic position not exceeding twice the modal average for that individual's chromosome. Fifteen individuals displayed evidence of high coverage at the position based on this criterion; as the cut-off was ≤ 14 high coverage individuals the position was filtered out.

Previous research had found this mutation occurs on one *Vgsc-1014F* bearing haplotype (Jones *et al.*, 2012a). To investigate whether this finding holds in Ag1000G, the filtered AR2 data set - which contained the *Vgsc-1575Y* variant - was used to create a network from exonic variation. The AR2 VGSC data produced a network similar in appearance to AR3, but with less reticulation between nodes, possibly due to an increased number of variants (Figure.4.3); there were 68 exonic variants in AR2 but only 60 in AR3. As with previous investigations into *Vgsc-1575Y* (Jones *et al*., 2012a), the resistant linked variant was found on a single haplotype, which bore the *Vgsc-1014F* mutation (Figure 4.3). The variant was present in countries where it had previously been detected (Burkina Faso and Cameroon) (Jones *et al.*, 2012a; Silva *et al*., 2014), but the network also revealed its presence in individuals from Guinea (Figure 4.3). The haplotype carrying the *Vgsc-1575Y* mutation was also the highest frequency tip node (Crandall and Templeton, 1993) in the star-shaped region of the network (Figure 4.3), which together with its presence across multiple countries and in agreement with analyses in Jones *et al*. (2012a) suggests a selective sweep acting upon it and strengthens the evidence for similar sweeps occurring on the other high frequency, multi-country, non-synonymous mutations found on *kdr* haplotypes such as the 1879 codon mutations. *Vgsc-1575Y* missing from the AR3 dataset highlights that conservative filtering regimes can remove 'true' genotypes calls along with errors and therefore raw data should not simply be discarded, but areas of interest should be investigated in detail. Genomic scans, for example for selection, may show signals due to LD but the actual causative variants may

not be present. Important SNPs, like this *kdr* mutation may have just failed filters but still have biological relevance.

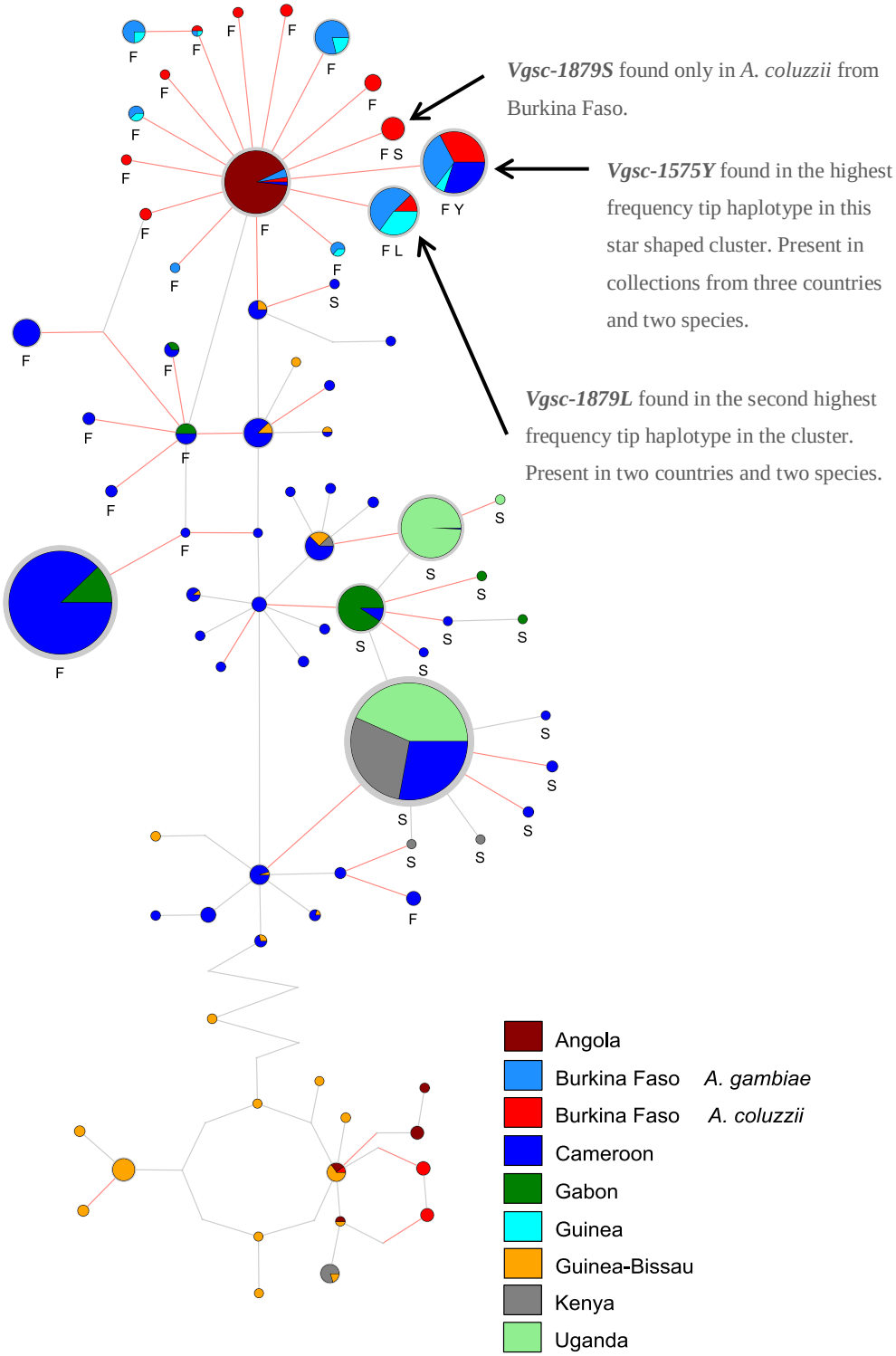


Figure 4.3. Haplotype parsimony network of exonic VGSC variation – AR2 data. Statistical parsimony network composed of phased haplotypes. Node size represents relative frequency of haplotypes, F = L1014F carrying haplotype, S = L1014S, F S = L1014F +

P1879S, F L = L1014F + P1879, F Y = L1014F + N1575Y. Red edges denote non-synonymous nucleotide changes, grey edges synonymous. Pie charts show proportion of haplotypes from country/species which make up each node.

4.4.5 Origins

Understanding the evolutionary origins of insecticide resistance is paramount if predictions are to be made about the effects of insecticidal vector control campaigns, the successes of which are essential in the fight against malaria (ffrench-Constant *et al*., 2004; World Health Organisation, 2012). The time taken for the seemingly inevitable evolution of resistance (Ranson *et al*., 2009) may depend on whether resistance conferring mutations are present in populations as standing variation (ffrench-Constant, 2007; Karasov *et al*., 2010), evolve *de novo* (Reimer *et al*., 2005) or enter through gene flow (Lynd *et al*., 2010). Where multiple resistance variants or haplotypes are found their efficacy of resistance may differ (*e.g.* Jones *et al*., 2012a) and therefore affect selection and gene flow. Here, investigation into the two best characterised point mutations, *Vgsc-1014F* and *S* conferring *kdr* resistance to the most important functional class of insecticides (Martinez-Torres *et* al., 1998; Ranson *et al*., 2000; van den Berg *et al*., 2012), revealed a complex evolutionary picture.

Previous estimations of the number of these *kdr* origins involving a network approach suggested four origins; however there are both limitations with data and issues with analyses (Pinto *et al*., 2007; Etang *et al*., 2009). Both of these studies, pre-WGS resources, used a small (~500bp) region from the intron preceding the exon containing the *kdr* mutations (reported as intron 1 in the papers, but actually intron 18 - AgamP4.2 gene set) which, despite many samples from across Africa, provided few variants to differentiate haplotypes, these variants were also uninformative about the functional evolution of the gene, being from a non-coding region. Interpretation of the networks in both these studies was also problematic, with coalescent theory regarding removal of ambiguous network connections (edges) being invoked in a non-neutral situation, invalidating assumptions of the method (Templeton, Crandall and Sing, 1992; Crandall and Templeton, 1993). The VGSC gene is an excellent example of a gene not undergoing neutral evolution (Lynd *et al*., 2010). If origins are defined, as in this analysis, as nodes (haplotypes) carrying *kdr* mutations, separated from other *kdr* nodes by nodes not carrying *kdr* mutations ('susceptible'), then reinterpretation of these previous studies' networks reduces the number of unambiguous origins to just two in the case of Pinto *et al*. (2007) and potentially just two in Etang *et al*. (2009). In the latter case

interpretation is hampered the inappropriate removal of "ambiguous" edges in the published paper.

Here, considering just the exonic variation in the VGSC across 1530 pan-African haplotypes (60 variants), four clear origins were visible for the two best characterised and widespread *kdr* mutations, two for *Vgsc-1014F* and two *Vgsc-1014S*, with these haplotypes geographically spanning the continent (Figure 4.1). However, as phylogenetic analysis of the VGSC shows high conservation across Insecta (Davies *et al*., 2007b), because of its role in the nervous system (Davies *et al*., 2007a) and because when investigated in other species, non-synonymous mutations were shown to be deleterious (Foster *et* al., 2003; Brito *et al*., 2013), it could be assumed the gene is under functional constraint. The addition of 46 non-coding variants from intron 18 to the network (in line with previous studies Pinto *et al.,* 2007; Etang *et al*., 2009; Santolamazza *et al*., 2015), revealed that convergence was indeed masking separate origins and actually eight origins were evident (three *Vgsc-1014F,* five *Vgsc-1014S*), six more than previously (unambiguously) suggested (compare Figure 4.1 with Figure 4.4) (Pinto *et al*., 2007; Etang *et al*., 2009). A clear example of this convergence was the high frequency *Vgsc*-1014S node, shared between Uganda, Cameroon and Kenya – suggesting long distance gene flow of over 2000km (Figure 1), was split by geography (Figure 4.4). Though this doesn't preclude a shared origin, but the 'susceptible' haplotypes that fall between the Cameroon and Uganda/Kenya resistance bearing nodes suggest separate origins are plausible.

Long range gene flow is still evident even after the addition of non-coding data however, and the different geographic origins revealed by the network do not necessarily have different *kdr* mutational origins, with one variant spanning the West African (Guinea Bissau, Guinea) and Central African samples sites (Cameroon, Angola) (Figure 4.4). Perhaps most striking, the star shape region of non-synonymous mutations revealed in the exonic network (Figure 4.1) was not changed by the additional non-coding variants and was still composed entirely of nodes separated by non-synonymous mutations (see discussion above and Figure 4.4). This suggests not only that evolution may be tinkering with the successful *Vgsc-1014F* haplotype to improve it (Jacob, 1977) but that this has happened so recently that no synonymous mutation (exonic or intronic) has had time to occur, possibly because these non-synonymous mutations confer a fitness advantage and their single haplotypes are currently sweeping through populations (see discussion above).

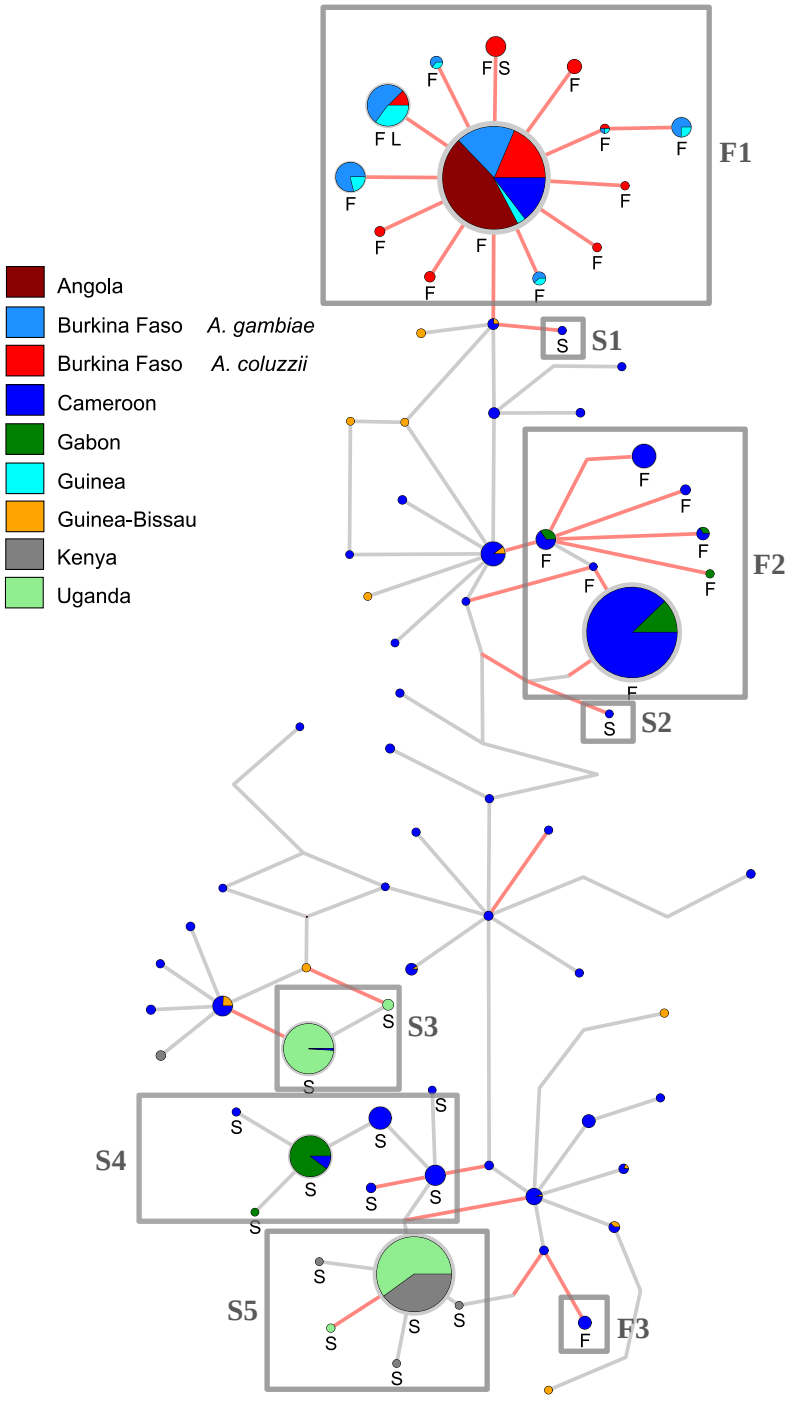


Figure 4.4. Haplotype network of exonic plus intron 18 VGSC variation – "kdr" region. Statistical parsimony network composed of phased AR3 haplotypes. Node size represents relative frequency of haplotypes, F = L1014F carrying haplotype, S = L1014S, F L = L1014F + P1879L, F S = L1014F + P1879S. Red edges denote non-synonymous nucleotide changes, grey edges exonic synonymous or intronic variants. Pie charts show proportion of haplotypes from country/species which make up each node and grey boxes highlight potential *kdr* mutation origins: where haplotypes carrying these mutations are separated from others by non-*kdr* ('susceptible') nodes. For full network see Appendix 4.6.5.

Additional variants from the AR2 data set increased the number of *Vgsc-1014F* origins from three to four and demonstrated that these values are only estimates based on the amount of data available. Further sampling, different filtering and additional populations included in future phases of the Ag1000G project will likely reveal even more potential origins of resistance. Numbers of origins are only estimates, as deeper sampling could reveal new nodes linking currently separated network origins (*i.e.* just diverse clades not new origins) and because recombination, although probably very rare and therefore unlikely to be driving network topography (Pinto *et al*., 2007), could be generating apparently novel resistance bearing nodes.

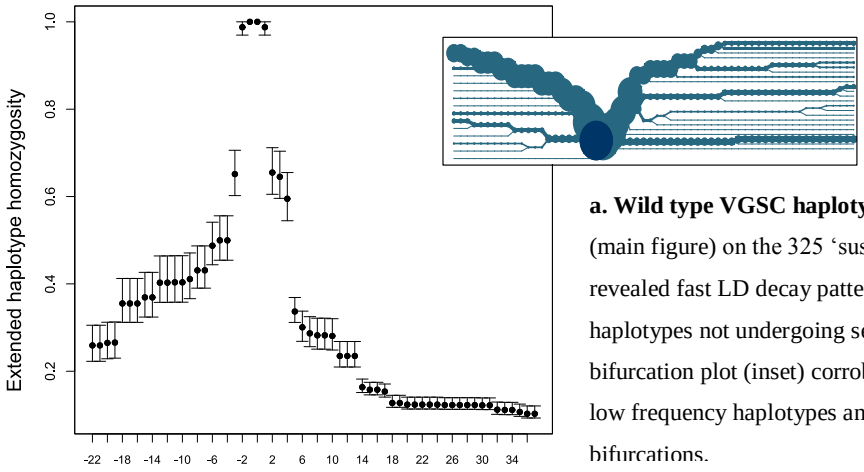
4.4.6 Extended haplotype homozygosity

EEH analysis was conducted on AR3 VGSC haplotypes, split into three groups based on *kdr* mutation (*Vgsc-1014S*, *Vgsc-1014F* and wild-type *Vgsc-1014L*), to reveal LD decay either side of the *kdr* mutation codon in the VGSC. Like the network analysis, EHH also suggests multiple origins of both *kdr* mutations with LD decay driven by several high frequency haplotypes in both centromeric and particularly telomeric directions shown by bifurcation plots (Figure 4.5b-c); this contrasts with the fast wild type LD decay around the 1014 codon which is driven by many lower frequency haplotypes (Figure 4.5a). A more general feature of the VGSC revealed by EHH analysis is a centromeric slowing in LD decay. In each grouping (*Vgsc-1014S*, *Vgsc-1014F* and wild-type *Vgsc-1014L*), LD does not decay as quickly over similar numbers of SNPs in the centromeric compared to the telomeric direction and there are also fewer bifurcations found in the centromeric direction (Figure 4.5). The centromere may be physically reducing recombination rates, making LD decay more slowly in proximity to it (Stump *et al*., 2005; Carneiro, Ferrand and Nachman, 2008).

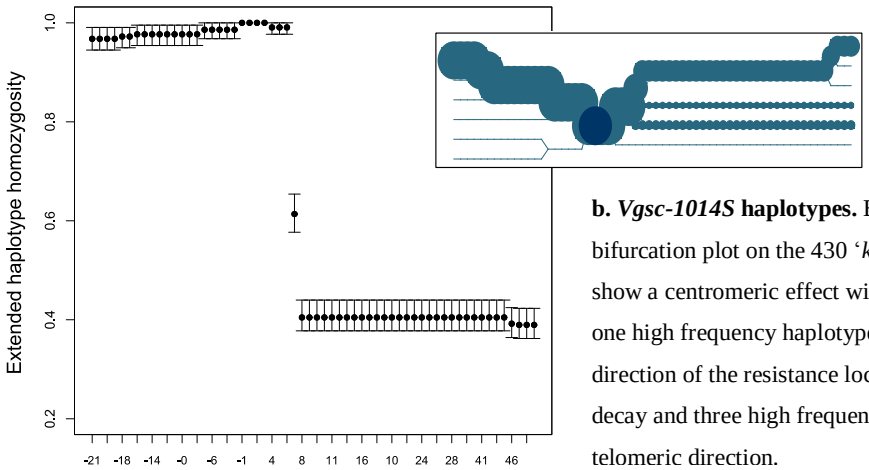
Differences between *Vgsc-1014F* and *Vgsc-1014S* results are striking, with slower LD decay and fewer (higher frequency) haplotypes bifurcations found across '*S*' haplotypes than '*F*' (Figure 4.5). One possible explanation for difference between the *kdr* mutations is the age of them. It has been suggested that *Vgsc-1014S* may be the older of these two *kdr* mutations, being selected by DDT (Lynd *et* al., 2010), the earliest insecticide widely adopted (Davies *et al*., 2007a); the selective pressure on *Vgsc-1014F* coming from pyrethroid insecticides which have only come into common use more recently (Davies *et al*., 2007a). This fits with the hypothesis that the excess of non-synonymous mutations found on the *Vgsc-1014F*

haplotypes are compensatory for the potential deleterious effects of the resistance mutation (Foster *et* al., 2003; Brito *et al*., 2013). The high frequency and tip placement of some of these non-synonymous carrying haplotypes in the network analyses suggests they are undergoing selective sweeps overlaid on top of the sweeps driven by the *kdr* mutation (Figure 4.1 and see Jones *et al*., 2012a). Potentially older, the *Vgsc-1014S* mutations may have had time to evolve compensatory mutations and for a few fitter haplotypes to sweep to high frequency and, at time of sampling, be the only haplotypes detectable. A study which performed a similar EHH analysis on a more localised and smaller data set collected between 1999-2005 adds further weight to this hypothesis as faster LD decay/more bifurcations were found in *Vgsc-1014S* haplotypes than in *Vgsc-1014F* (Lynd *et al*., 2010), the opposite to what we found in our analysis conducted on samples collected more recently (Gabon – 2000, all other populations 2009-2012). The cause of these contrasting results could therefore be due to a very recent evolution of optimised *Vgsc-1014F* haplotypes sweeping across Africa (with the effect of reducing LD decay) with the faster LD decay/bifurcation found in *Vgsc-1014F* haplotypes being driven by attempts to ameliorate deleterious effects (in certain environments) of carrying a more recently selected *Vgsc-1014F kdr* mutation.

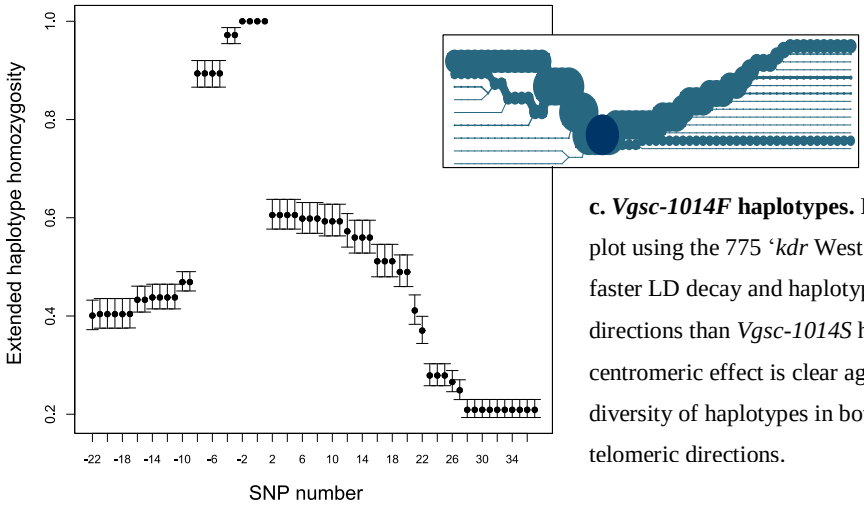
It is clear that the same *kdr* point mutations in the VGSC gene have evolved multiple times, the gene undergoing multiple global soft sweeps with several origins being found within populations. Insecticide pressure and gene flow has also spread haplotypes across Africa with much haplotype sharing between even distant countries. However, the quest for the true number of origins is trivial, from both a medical and evolutionary perspective. The important message is that when faced with strong anthropogenic insecticide pressure, the VGSC undergoes repeated convergent evolution, generating repeated local selective sweeps. Although these present a classic hard sweep signature on a local scale (Lynd *et al*., 2010), they are soft sweeps on a pan-African one (Messer and Petrov, 2013). Selective sweeps overlaid on top of 'resistant' haplotypes also reveal that VGSC adaptation doesn't stop once *kdr* mutations are at high frequency within populations.



a. Wild type VGSC haplotypes. EHH analysis (main figure) on the 325 'susceptible' haplotypes revealed fast LD decay pattern expected from haplotypes not undergoing selective sweep. The bifurcation plot (inset) corroborates this with many low frequency haplotypes and manifold bifurcations.



b. *Vgsc-1014S* haplotypes. Both EHH and bifurcation plot on the 430 '*kdr* East' haplotypes show a centromeric effect with slow LD decay and one high frequency haplotype in a centromeric direction of the resistance locus but faster LD decay and three high frequency haplotypes in the telomeric direction.



c. *Vgsc-1014F* haplotypes. EHH and bifurcation plot using the 775 '*kdr* West' haplotypes show faster LD decay and haplotype bifurcation in both directions than *Vgsc-1014S* haplotypes. A centromeric effect is clear again, through there is a diversity of haplotypes in both centromeric and telomeric directions.

Figure 4.5. EHH analysis showing VGSC exonic LD decay with increasing distance from the core and (inset) bifurcation plots showing recombination patterns for these SNPs in wild type, Vgsc-1014S and Vgsc-1014F haplotypes. Core defined in EHH figures as the origin on the X axis and in bifurcation plots as the darker blue circle. Bars in EHH figures represent 95% CIs calculated via bootstrapping. Circle size in bifurcation plot represents frequency of haplotypes.

4.4.7 Cameroonian hyper diversity

Gene flow is clearly a major mechanism involved in the evolution of resistance to insecticide, however, most countries appear to only host one resistance origin with Uganda (two *Vgsc-1014F)* and Gabon (one *Vgsc-1014F*, one *Vgsc-1014S*) having two (Table 4.1). This finding fits a model of strong insecticide pressure causing selective sweeps on a single or a small number of fitness conferring resistance haplotypes within a short space of time (Lynd *et al*., 2010). The star shaped network features, particularly visible in *Vgsc-1014F* haplotypes, also suggest recent sweeps on single haplotypes (Figure 4.1). In contrast to this, the network reveals that Cameroon supports, putatively, seven of the eight resistance origins (Figure 4.4; Table 4.1). Whilst Cameroon has the largest sample size in Ag1000G and three collection sites, the different origins are not localised to collection sites but are distributed across them (Appendix 4.6.7). Therefore, it appears that something else is driving the patterns of VGSC diversity found there. Perhaps historically the country had strong population subdivision or a particularly heterogeneous insecticide use landscape that may have allowed evolution of many different *kdr* resistance origins. How then did these resistance mutations seed much of the *kdr* resistance found across the continent? These are currently only speculative hypotheses, however, it is clear that understanding the evolution of resistance in Cameroon is vital to understanding the phenomenon on an Africa wide scale.

Table 4.1. Haplotype frequencies and countries binned by *kdr* mutation origin.
Table shows frequencies of all 1530 haplotypes (765 individuals) for species and country of collection, potential *kdr* origin (**F1-3, S1-5**) and 'susceptible' (**sus**) haplotypes (see Figure 4.3). For all countries without species labelling the mosquitoes were *A. gambiae*.

Origin	Burkina Faso - *A. coluzzii*	Burkina Faso - *A. gambiae*	Uganda	Kenya	Angola	Gabon	Guinea	Guinea Bissau	Cameroon	Total haplotypes
F1	117	162	0	0	103	0	62	0	33	477
F2	0	0	0	0	0	40	0	0	245	285
F3	0	0	0	0	0	0	0	0	13	13
S1	0	0	0	0	0	0	0	0	2	2
S2	0	0	0	0	0	0	0	0	1	1
S3	0	0	108	0	0	0	0	0	1	109
S4	0	0	0	0	0	72	0	0	81	153
S5	0	0	98	67	0	0	0	0	0	165
sus	21	0	0	21	17	0	0	92	174	325

4.4.8 Susceptible haplotypes

Multiple origins of *kdr* mutations are suggested from the network - clusters of resistance mutation bearing haplotypes separated by 'susceptible' (wild type) haplotypes (Figure 4.1). However, the resistance haplotypes are more closely related to each other in contrast to the susceptible haplotypes which display much more variation, some clusters separated from their most closely related haplotypes by many divergent SNPs (Figure 4.1), EHH analysis and bifurcation plots also reveal the high diversity in 'susceptible' haplotypes relative to *kdr* carrying haplotypes (Figure 4.5). This wild type diversity is even more striking with intronic variation included in the network as divergence between these susceptible haplotypes is almost entirely composed of exonic synonymous and intronic variants (Figure 4.6). The fewest mutational steps between the most divergent resistant haplotypes was 16, compared to 37 steps for the most divergent susceptible haplotypes (Appendix 4.6.5). This contrast was even more extreme with the additional variation in the AR2 data set (Appendix 4.6.6). Guinea Bissau, a country without any VGSC resistance mutations in the Ag1000G phase 1 data was the collection location for the most distant susceptible haplotypes (Figure 4.6). With recent research by the 16 Genomes Consortium revealing massive inter-specific introgression within the genus (Fontaine *et al*., 2015; Neafsey *et al*., 2015), one hypothesis for this diversity is introgression of regions of genome containing the VGSC gene from other species.

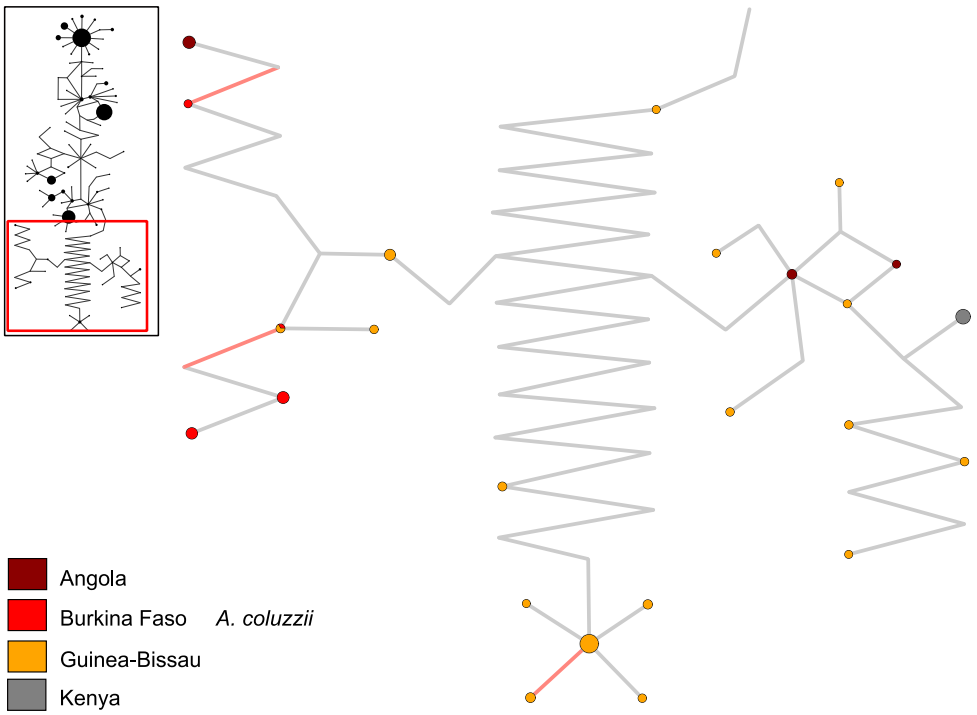


Figure 4.6. Haplotype network of exonic plus intron 18 VGSC exonic variation – "susceptible" region. Statistical parsimony network composed of phased AR3 haplotypes. Node size represents relative frequency of haplotypes. Red edges denote non-synonymous nucleotide changes, grey edges exonic synonymous or intronic variants. Pie charts show proportion of haplotypes from country/species which make up each node.

To test the hypothesis that introgression is driving susceptible haplotype diversity, a phylogenetic approach was taken. High frequency haplotypes representative of regions of the network (Appendix 4.6.2) were reconstructed and aligned to the same genomic region taken from the reference genomes of other *Anopheles* species, potential sources of introgression, particularly in Guinea Bissau (highly diverse VGSC haplotypes) where *A. arabiensis* and *A. melas* are found in sympatry with *A. gambiae* and *coluzzii* (Gordicho *et al*., 2014). *A. funestus* was included as a more divergent outgroup but dropped out of the analysis due to high sequence divergence from all other samples. The maximum likelihood phylogeny revealed that all the network haplotypes were more similar to each other and the *A. gambiae* and *A. coluzzii* reference genomes (for example a 70% consensus would collapse them into a single node with 99% bootstrap support) than to *A. arabiensis* or the more distant *A. melas* (Figure 4.7). The *A. coluzzii kdr* resistant haplotype (haplotype F1) falling with the *A. gambiae* clade was expected due to the introgression of resistance (Lynd *et al*., 2010). However, Burkina Faso *A. coluzzii* susceptible haplotype (haplotype wt2) also fell within this clade, rather than with the *A coluzzii* reference genome, perhaps highlighting that

sequencing wild mosquitoes for comparison would be optimal rather than the reference genomes used here.

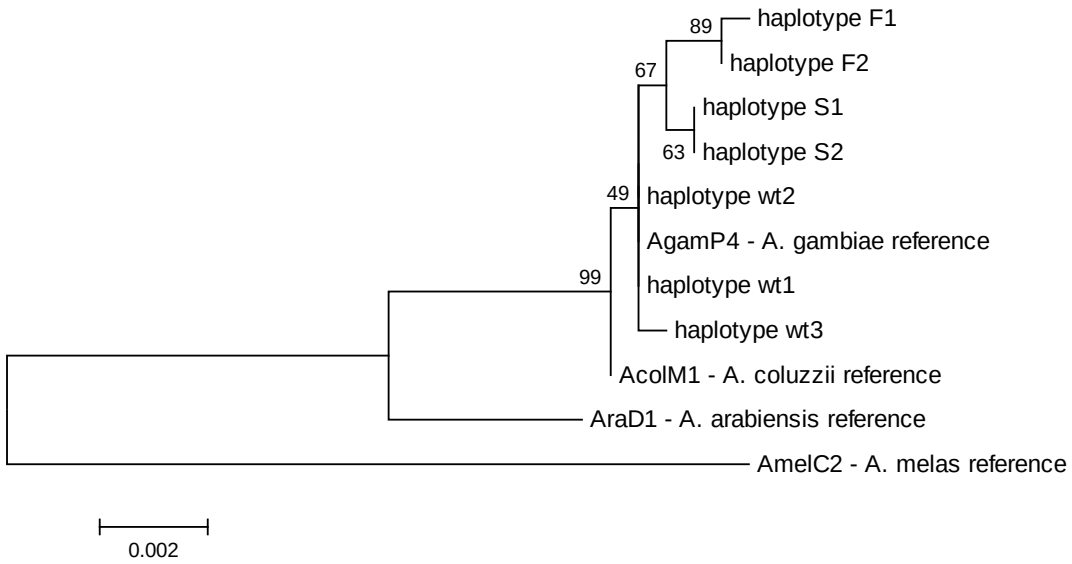


Figure 4.7. Anopheles phylogeny of VGSC intron 18 and exon 19 genomic region. A maximum likelihood phylogeny using a 2000 nucleotide sequence composed of VGSC exon 19 and the preceding region of intron 18. Four species' reference genomes were compared with seven haplotypes taken from phased AR3 data. 1000 bootstrap iterations were performed. wt*n* = wildtype haplotype *n*, S*n* = *Vgcs-1014S* haplotype *n*, F*n* = *Vgsc-1014F* haplotype *n*. Versions of reference genomes used are denoted by code *e.g.* AmelC2.

As high levels of the variation found is non-synonymous and therefore less affected by purifying selection, an alternative hypothesis to inter-specific introgression is simply that these levels of neutral diversity are expected and only seem high compared to the selectively swept and thus variation depauperate resistance bearing haplotypes. Whether this is indeed expected diversity warrants further investigation, for example by simulations of evolving populations (Ewing and Hermisson, 2010; Ohashi, Naki and Tsuchiya, 2011), however understanding the recombination landscape of *A. gambiae* through the development of a high resolution recombination map would be highly advantageous in this endeavour (Hamblin and Aquadro, 1996). The relative diversity of the susceptible haplotypes also highlights the *kdr* origins evolving on similar haplotypes. With findings of *kdr* mutation fitness costs in the housefly *Musca domestica* (loss of larval temperature gradient preference) and dengue vector mosquitoe *Aedes Aegypti* (slowed development, reduced eggs) (Foster *et* al., 2003; Brito *et*

al., 2013), it may be that only some haplotypes can support the conformational change of the *kdr* mutations without lethality. However, more functional work, and an understanding of the deleterious effects of *kdr* is required.

4.4.9 Future work

The network analysis has revealed many interesting and hitherto unknown patterns of evolution in the *A. gambiae* VGSC gene, generating a plethora a new hypotheses to test.

- One obvious avenue for further investigation is into the non-synonymous mutations found on *kdr* resistance mutation bearing haplotypes; do these functional changes increase resistance or are they compensatory for the fitness costs of *kdr*? Presently the 3D protein models of the VGSC gene are not up to this task, these could be extended to encompass the whole gene to allow an *in silico* functional assessment of the variant effects (Davies *et al*., 2007b).
- *In vitro* investigations could be employed, inserting these mutations into *Xenopus* oocytes to test the function of the channel in the presence and absence of insecticides (Burton *et al*., 2011).
- *In vivo*, lab strains containing different combinations of non-synonymous mutations could also be created to allow tests of fitness under different simulated ecological scenarios. Bioassays and haplotypic association test should be repeated for the *Vgsc-1879* mutations to determine if they confer additional insecticide resistance.
- Network and EHH analyses suggest multiple selective sweeps occurring on *kdr* haplotypes, possibly increasing resistance or ameliorating fitness costs. *Vgsc-1014S* haplotypes appear to have lost diversity since results published in a 2010 study, potentially through optimal haplotypes evolving and sweeping across already resistant populations. Older samples could be compared with those used in this study to identify potential SNPs, allowing the selective sweeps to be tracked over time and space, while allowing predictions to be made about the evolution of *Vgsc-1014F*.
- Alongside questions about the function of the VGSC, questions about the evolutionary/population history and population genetics of the gene have also been raised. For example, Cameroon supports *kdr* bearing haplotypes from a number of different origins; the demographic history may help explain the many potential origins of *kdr* mutations found there *e.g.* is the diversity driven by secondary contact of subdivided populations?
- A coalescent simulation approach could be used to date and evaluate selection on the *kdr* mutations (Ohashi, Naki and Tsuchiya, 2011), though this would require a larger

sample of susceptible *A. gambiae* haplotypes from a population with resistance mutations than was available in phase 1 of Ag1000G.

- To understand the nucleotide variation found in the VGSC, purifying and background selection could be investigated, thought a high resolution recombination map, currently unavailable for the species, would be necessary (Charlesworth, 1996).

4.4.10 Conclusions

The network analysis has demonstrated that a simple technique can be powerful, allowing large and complex data sets to be visualised in a manner from which many inferences about evolutionary processes can be gleaned for further investigation. The network allows estimation of multiple origins for *kdr* mutations, but also shows that gene flow is a major force in the evolution of resistance with some haplotypes shared over large tracts of Africa. Striking VGSC genetic complexity was discovered, both synonymous within susceptible haplotype and particularly the non-synonymous protein altering variation found on *kdr* containing haplotypes, suggesting selection on compensatory or additive insecticide resistance effects. The former revealing how diverse the gene was prior to the selective sweeps driven by the insecticide resistance conferring *kdr* mutation. The latter showing that the standard practice of genotyping individuals at just three non-synonymous insecticide resistance linked mutations (*Vgsc-1014F*, *Vgsc-1014S* and *Vgsc-1575Y*) may be underestimating medically important functional differences with respect to malaria vector resistance or fitness phenotypes. The challenge now must be to understand this abundance of non-synonymous mutations as it is clear that current molecular genotyping efforts are failing to realise this diversity.

4.5 Acknowledgements

The Ag1000G Consortium collected, sequenced and performed quality control on all samples and discussion of findings with the Ag1000G Analysis Team greatly enriched this chapter. Emyr (T. G. E.) Davies (Rothamstead Research) was responsible for 3D protein modelling. Emily Rippon carried out cloning of ambiguous double heterozygote individuals for the 1879 TaqMan assays validation.

4.6 Appendix

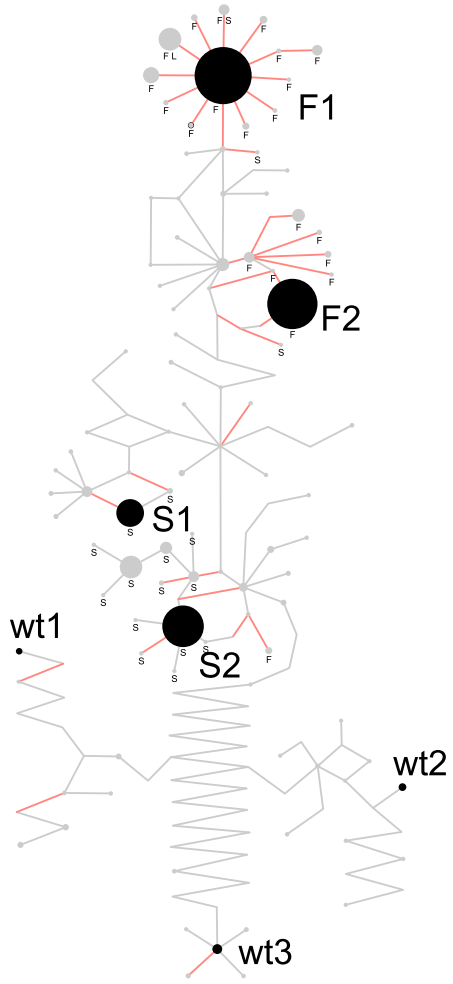
Appendix 4.6.1 –V*gsc*-1879 mutation genomic region consensus sequence for TaqMan design.

Variation from phased haplotype analysis collated and added to consensus sequence as IUPAC ambiguity codes to allow variation sensitive design of TaqMan primers and probes.

>1879_TAQMAN_2L_2430657_2431079

YTGTCTTCCACCAGACAATGATAAGGGCTATCCGGGAAATTGTGGTTCATCAACAATTGGCATAACGTACTTATT
GGCGTATCTTGTAATAAGTTTCCTTATCGTTATTAACATGTACATTGCTGTTATYCTCGAAAACTACTCGCAAGCT
ACGGAAGATRTTCAAGAAGGCTTAACTGATGACGATTATGATATGTACTAYGAAAYATGGCAGCAATTCGATYY
TGACGGWACACAATACGTTCGATATGATCAGCTRTCAGACTTTTTGGATGTGCTGGAACCGCCTCTACAGATTCA
TAAACCAAATCGTTATAAGATTATTTCGATGGATATTCCGATATGCCGYGGAGATATGATGTYCTGTGTCGATAT
TCTAGATGCACTAACGAAAGATTTTTTTGYTAGAAAAGGAAATCCTAY

Appendix 4.6.2 – representative haplotypes



Appendix 4.6.2. Representative haplotypes for phylogenetic analysis. Statistical parsimony network composed of phased AR3 exons plus intro 18 variant haplotypes. Node size represents relative frequency of haplotypes, F = L1014F carrying haplotype, S = L1014S, F S = L1014F + P1879S, F L = L1014F + P1879. Black nodes and bold text denote chosen haplotypes.

Appendix 4.6.3 VGSC resistance linked mutations

Rinkevich et al 2013. Mapped from house fly VGSC (genbank:X96668) to A. gambiae VGSC

#Table 1. Detected in >1 species and functionally characterised

V410? In all VGSC RA, RB and RC this is position 402 - AGAMP4.2 codon = GTA 2L:2391228-2391230

M827? In all VGSC RA, RB and RC this is position 808 - AGAMP4.2 codon = ATG 2L:2417030-2417032

M918? In all VGSC RA, RB and RC this is position 899 - AGAMP4.2 codon = ATG 2L:2417694-2417696

L925? In all VGSC RA, RB and RC this is position 906 - AGAMP4.2 codon = CTT 2L:2417715-2417717

T929? In all VGSC RA, RB and RC this is position 910 - AGAMP4.2 codon = ACC 2L:2417727-2417729

L932? In all VGSC RA, RB and RC this is position 913 - AGAMP4.2 codon = ACC 2L:2417736-2417738

I1011? In all VGSC RA, RB and RC this is position 992 - AGAMP4.2 codon = ATA 2L:2422641-2422643

L1014? In all VGSC RA, RB and RC this is position 995 - AGAMP4.2 codon = TTA 2L:2422650-2422652

L1016? In all VGSC RA, RB and RC this is position 997 - AGAMP4.2 codon = GTG 2L:2422713-2422715

L1020? In all VGSC RA, RB and RC this is position 1001 - AGAMP4.2 codon = TTC 2L:2422725-2422727

L1534? VGSC RA this is position 1529, RB and RC position 1503 - AGAMP4.2 codon = TTT 2L:2429622-2429624

L1538? VGSC RA this is position 1533, RB and RC position 1507 - AGAMP4.2 codon = TTC 2L:2429634-2429636

D1549? VGSC RA this is position 1544, RB and RC position 1518AGAMP4.2 codon = GAC 2L:2429667-2429669

D1553? VGSC RA this is position 1548, RB and RC position 1522 - AGAMP4.2 codon = GAA 2L:2429679-2429681

Table 2. Single species

I254? In all VGSC RA, RB and RC this is position 247 - AGAMP4.2 codon = ATA 2L:2390155-2390157

E435? In all VGSC RA, RB and RC this is position 427 - AGAMP4.2 codon = GAG 2:2391303-2391305

C785? In all VGSC RA, RB and RC this is position 766 - AGAMP4.2 codon = TGT 2:2416904-2416906

Q945? In all VGSC RA, RB and RC this is position 926 - AGAMP4.2 codon = CAA 2L:2417775-2417777

Q979? In all VGSC RA, RB and RC this is position 960 - AGAMP4.2 codon = TTC 2L:2422545-2422547

Q989? In all VGSC RA, RB and RC this is position 970 - AGAMP4.2 codon = TCA 2L:2422575-2422577

Q1010? In all VGSC RA, RB and RC this is position 991 - AGAMP4.2 codon = GTG 2L:2422638-2422640

Q1013? In all VGSC RA, RB and RC this is position 994 - AGAMP4.2 codon = AAT 2L:2422647-2422649

Q1020? In all VGSC RA, RB and RC this is position 1005 - AGAMP4.2 codon = CTT 2L:2422737-2422739

Q1060? Aligns poorly to housefly, not possible to determine position.

Q1215? Aligns poorly to housefly, not possible to determine position.

N1410? VGSC RA this is position 1402, RB and RC position 1376 - AGAMP4.2 codon = GCG 2L:2429109-2429111

N1494? VGSC RA this is position 1486, RB and RC position 1460 - AGAMP4.2 codon = GCC 2L:2429423-2429425

M1524? VGSC RA this is position 1519, RB and RC position 1493 - AGAMP4.2 codon = ATG 2L:2429592-2429594

M1528? VGSC RA this is position 1523, RB and RC position 1497 - AGAMP4.2 codon = TTT 2L:2429604-2429606

N1575? VGSC RA this is position 1570, RB and RC position 1544 - AGAMP4.2 codon = AAT 2L:2429745-2429747

M1752? VGSC RA this is position 1747, RB and RC position 1721 - AGAMP4.2 codon = ATT 2L:2430427-2430429

V1823? VGSC RA this is position I1818, RB and RC position I1792 - AGAMP4.2 codon = ATT 2L:2430712-2430714

P1879? VGSC RA this is position 1874, RB and RC position 1848 - AGAMP4.2 codon = CCT 2L:2430880-2430882

Table 3. Confirmed in Xenopus

L933? In all VGSC RA, RB and RC this is position 914 - AGAMP4.2 codon = CTT 2L:2417739-2417741

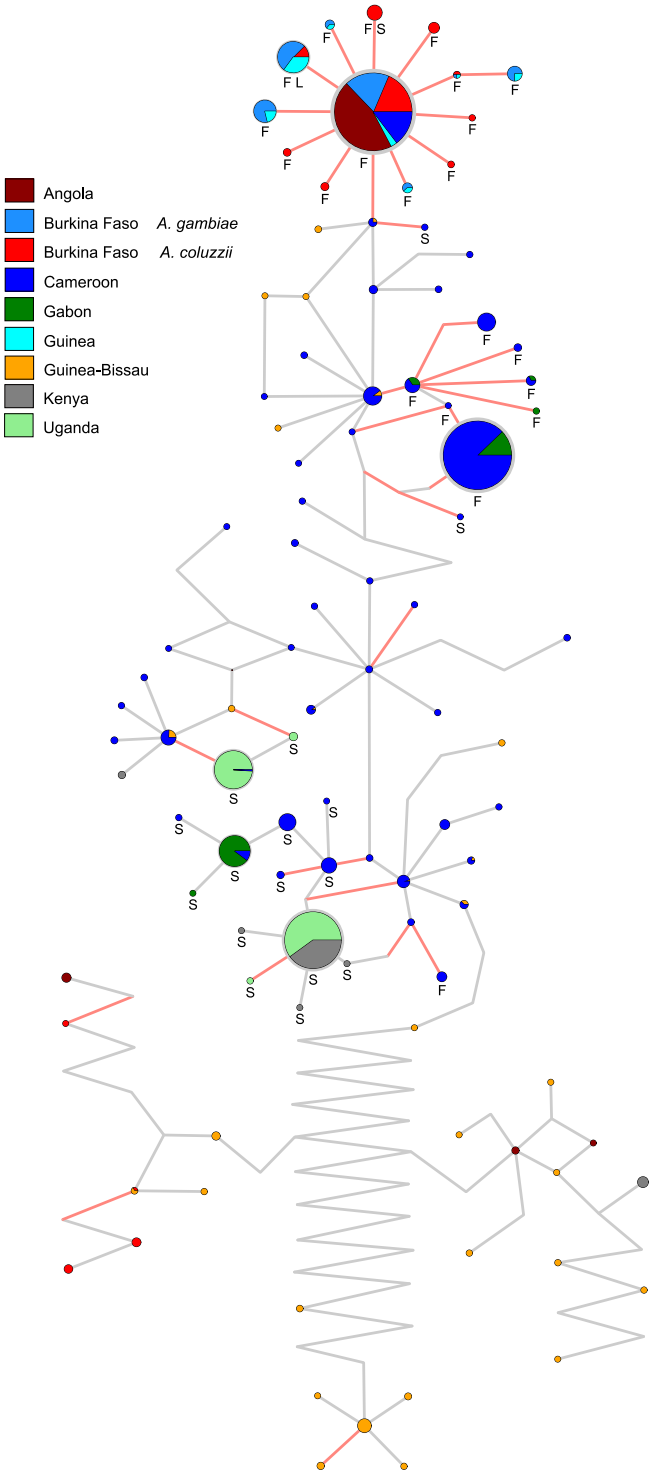
I936? In all VGSC RA, RB and RC this is position 917 - AGAMP4.2 codon = CTT 2L:2417748-2417750

D1596? VGSC RA this is position 1591, RB and RC position 1565 - AGAMP4.2 codon = CCT 2L:2429878-2429880

Appendix 4.6.4 – Taqman QC

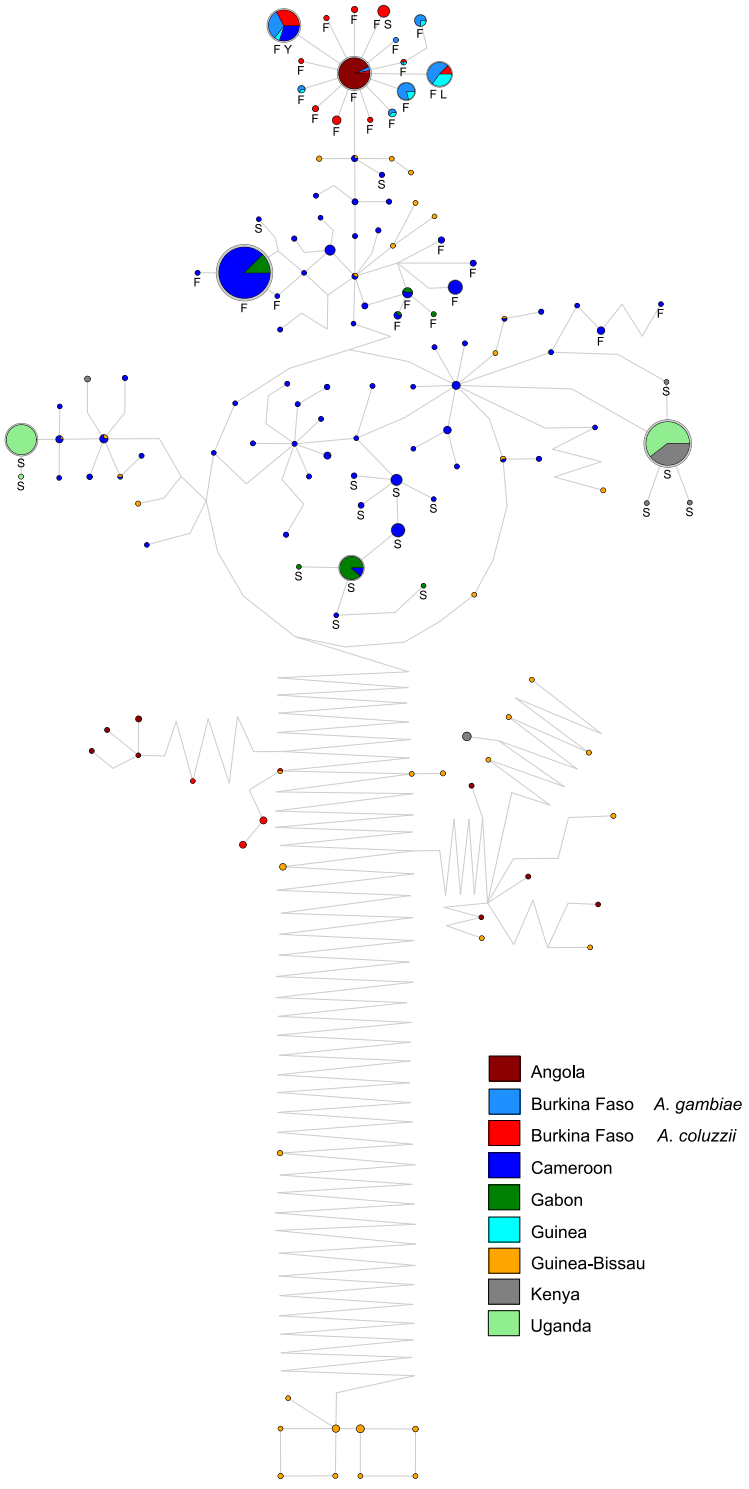
PCR and sequencing of Tiassalé colony female individuals revealed all genotypes estimated by TaqMan were correct, except for individuals except individuals that appeared heterozygous for both *Vgsc-1879L* and *Vgsc-P1879S* mutations (Genbank accession: KT290212 - KT290220). There was a possibility that rather than the 1879 codons in the two homologous chromosomes being TCT (*L*) and CTT (*S*), the same Taqman genotypes could result from a recombination event or mutation bringing both 1879 mutations together on the same chromosome (TTT – *Vgsc-1879F*) in concert with a wildtype codon on the other chromosome (CCT - *Vgsc-1879P* - wildtype). To elucidate these genotypes cloning and sequencing was utilised which found, in 12 'double heterozygote' females, no evidence for both mutations falling on a single haplotype; all codons were either TCT (P1879S) or CTT (P1979L) (Genbank accession: KT290221 – KT290232).

Appendix 4.6.5 – whole AR3 network



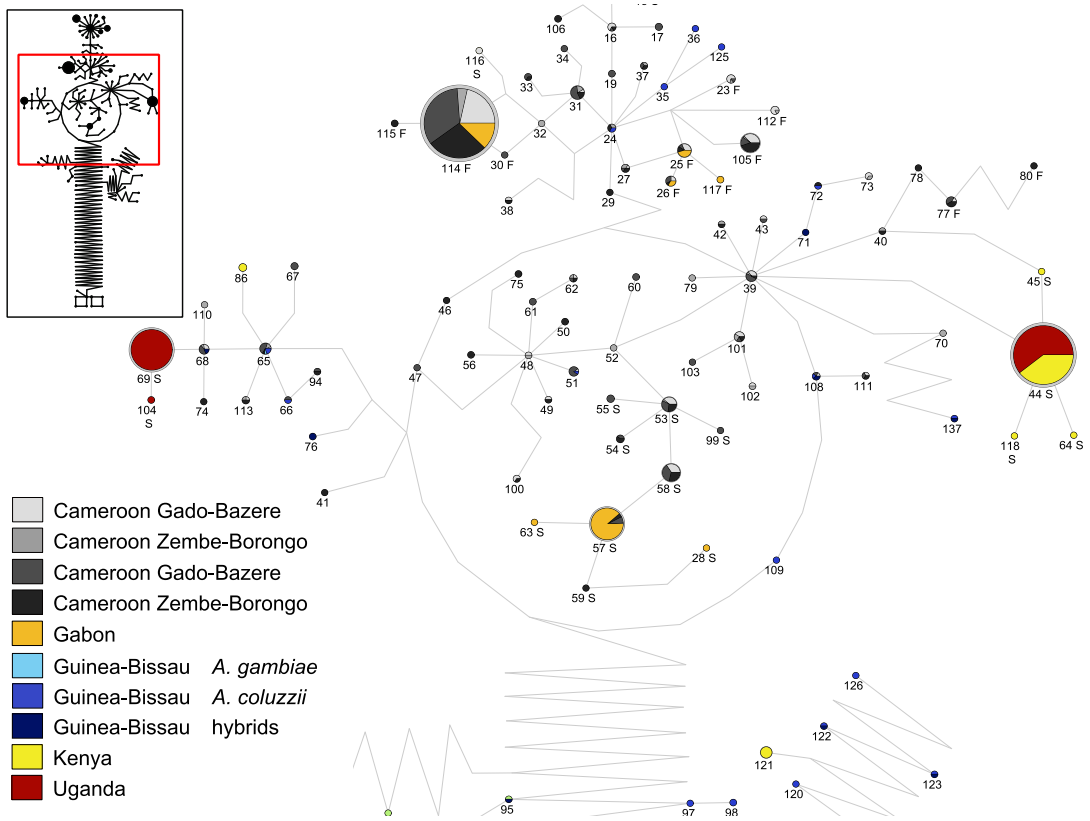
Appendix 4.6.5. Haplotype parsimony network of exonic and intron 18 VGSC variation – AR3. Statistical parsimony network composed of phased AR2 haplotypes. Node size represents relative frequency of haplotypes, F = L1014F carrying haplotype, S = L1014S, F S = L1014F + P1879S, F L = L1014F + P1879, F Y = L1014F + N1575Y. Red edges denote non-synonymous nucleotide changes, grey edges synonymous. Pie charts show proportion of haplotypes from country/species which make up each node.

Appendix 4.6.6 – whole AR2 network



Appendix 4.6.6. Haplotype parsimony network of exons plus intron 18 VGSC variation– AR2. Statistical parsimony network composed of phased AR2 haplotypes. Node size represents relative frequency of haplotypes, F = L1014F carrying haplotype, S = L1014S, F S = L1014F + P1879S, F L = L1014F + P1879, F Y = L1014F + N1575Y. Pie charts show proportion of haplotypes from country/species which make up each node.

Appendix 4.6.7- Cameroon split by collection site



Appendix 4.6.7. Haplotype parsimony network of exons plus intron 18 VGSC variation– AR2 'Vgsc-1014S region'– Cameroon split. Statistical parsimony network composed of phased AR2 haplotypes. Node size represents relative frequency of haplotypes. Pie charts show proportion of haplotypes from country/species/collection site which make up each node. This example region shows how haplotypes are shared across Cameroon's four collection sites.

Chapter 5

Anthropogenic adaption in *Anopheles arabiensis*: identifying insecticide resistance candidates using pooled sequencing for genome-wide association

5.1 Abstract

Malaria vector control programmes have seen great successes in recent years; however the gains made in the fight against the disease are vulnerable. The feeding and resting behaviours of mosquitoes are often what expose them to insecticides; ITN and IRS campaigns are most effective for endophilic and endophagic vectors species such as *A. gambiae*. For *A. arabiensis*, a species with more variation in behaviour, often feeding and resting outside, these methods may have reduced efficacy. With suggestions of species composition shifting toward *A. arabiensis* through climate change, human movement and loss of other species due to malaria control, its prominence as a vector may increase; therefore an understanding of insecticide resistance in this species is timely and apposite. Recent advances in the genomic resources publically available for *A. arabiensis* enabled the first pool-seq GWAS in *Anopheles*. Using this approach the genome-wide allele frequencies of over 1000 resistant and susceptible mosquitoes were compared in pools, to identify candidates driving resistance phenotypes. By using intra- and inter-population comparisons from mainland Tanzania and from Zanzibar, with replication, several regions of the genome were found associated with pyrethroid resistance, including one on the 2R chromosome arm which was found in all resistant *versus* susceptible comparisons. The ~225kb 2R candidate region coincided with cluster of CYP450 genes (metabolisers of xenobiotics) from the 6 sub-family, including *Cyp6p4*, previously shown to metabolise the pyrethroids and found upregulated in resistant *A. arabiensis*. Investigation of nucleotide variation in this region found no single resistance associated SNP across all comparisons, possibly suggesting multiple soft selective sweeps driving resistance. Results demonstrate that by using pool-seq

to generate statistical power and reduce cost, association studies can generate confident phenotype-genomic region associations in *A. arabiensis*.

5.2 Introduction

5.2.1 Malaria vectors

Recent years have seen great headway made in the fight against malaria, with vector control playing a large part in a drop of over 50% in deaths from the parasite in sub-Saharan Africa since the year 2000 (World Health Organisation, 2011; World Health Organisation, 2014). However, research suggests these successes in malaria control may be vulnerable to shifts in vector species composition due to species-dependent efficacy of approaches (Kitau *et al*., 2012). For example the effectiveness of insecticide treated bednets (ITN) and indoor residual spraying (IRS) campaigns rely on vector feeding and resting behaviours, known to differ between common vector species. *A. gambiae* and *A. funestus* are generally nocturnal, endophilic, endophagic and anthrophilic and therefore are excellent targets for ITN and IRS (Pates and Curtis, 2005; Lyimo and Ferguson, 2009; Kitau *et al*., 2012) but *A. arabiensis* are known to have more diverse behaviour, feeding and resting both indoors and out (Sinka *et al*., 2010; Gordicho *et al*., 2014). In addition, *Anopheles arabiensis* exposed to insecticide treated bednets in Tanzania had reduced mortality compared to other local vectors, *A. gambiae* and *A. funestus*, and the authors of that study concluded that *A. arabiensis* may become the main transmitter of malaria transmission in the future (Kitau *et al*., 2012).

Though *A. arabiensis* is not a neglected vector species (*e.g.* see references above), research emphasis has been placed upon the major malaria vector sibling species *A. gambiae* and *A. coluzzii* in part due to excellent genetic resources and long standing, well annotated reference genome (Holt *et al*., 2002; Giraldo-Calderón *et al*., 2014). A Google Scholar search for "Anopheles gambiae" since 2011 returns ~15,200 hits compared with "Anopheles arabiensis" only returning ~2510 (searched 01th October 2015). "Anopheles gambiae s.l." only returns ~1180 hits so is not drastically inflating the *A. gambiae* hits (searched 01th October 2015). However, the medical importance of *A.arabiensis* as a vector may be on the increase, not only through a behavioural avoidance of control efforts (Bayoh *et al*., 2010; Kitau *et al*., 2012), but also through climate change. A loss of *A. gambiae* and *A. funestus* in Tanzania is thought to have been influenced by declining precipitation (Meyrowitsch *et al*., 2011), something projected to increase across most tropical and sub-tropical regions (Parry

et al., 2007). With *A. arabiensis* seemingly favouring drier environments (Sinka *et al.*, 2010), they may become dominant vectors.

Changes in the way humans live in sub-Saharan Africa are also increasing the importance of *A. arabiensis*. Djogbénou *et al*. (2008a) found that it outnumbered *A. gambiae* in a Bobo-Dioulasso, a city in Burkina Faso, and research suggests *A. arabiensis* are able to survive in polluted larval habitats generally thought as unsuitable for other *Anopheles* (Chinery, 1984), possibly through upregulated detoxification genes and point mutations (Jones *et al*., 2012a). It should be noted, however that there is also evidence of adaptation to urban environments by another vector, *A. coluzzii* (Cassone *et al*., 2014). With 56% of Africans predicted to be living in urban areas by 2050 (World Urbanization Prospects, 2014), it is clear that a greater understanding of resistance to xenobiotics in *A. arabiensis* will aid future vector control, both through understanding and predicting species shifts and in overcoming insecticide resistance. Fortunately the *Anopheles* 16 Genomes Project (Neafsey *et al*., 2015) has catalysed genomic research into species other than *A. gambiae* with improvement of reference genomes, including annotation of *A. arabiensis* genes.

5.2.2 Genome wide association studies

The improvement in, and accessibility of, *A. arabiensis* genomic resources now allow powerful techniques to be employed that can uncover loci driving resistance to xenobiotics. Genome wide association studies (GWAS) use dense genetic markers covering a large proportion of the genome, often from SNP chips or whole genome sequencing (WGS), to map the association of a genotype to a trait under examination (McCarthy *et al*., 2008). Initially the technique was developed to understand the genetic bases of human diseases, with case/control studies successful in locating genes and SNPs involved wide ranging illnesses such as cancers (Eeles *et al*., 2008; Hunter *et al*., 2007) and diabetes (Gudmundsson *et al*., 2007). However, increases in genomic resources has seen the technique applied across model organisms, including *Drosophila* (Turner, Miller and Cochrane, 2013) and mice (Flint and Eskin, 2012), and non-model organisms of agricultural importance such as cattle (Olsen *et al*., 2011).

Prior to the advent of GWAS, quantitative trait loci (QTL) mapping was the technique generally used to locate regions of the genome associated with a trait. However, this requires

several generations of crosses to be made between the treatment groups and often necessitates inbred lines to be made prior to crossing. Not only does this make QTL mapping difficult to perform on wild populations (Witzig *et al*., 2013), but the resolution is limited by the amount of recombination occurring in the crosses (Korte and Farlow, 2013). GWAS allows samples differing in traits of interest to be taken from natural population with resolution limited only by ancestral recombination and linkage disequilibrium (LD). However, conducting a GWAS is not without its own difficulties; Large numbers of individuals are often necessary to generate statistical power, *i.e.* for candidates to pass significance thresholds, particularly if effect sizes are small (Turner, Miller and Cochrane, 2013). Though the cost of sequencing has dropped drastically over recent years (Baker, 2010), with the requirements for hundreds to hundreds of thousands of individual genomes to produce statistical power (Manolio *et al*., 2009; Park *et al*., 2010), the costs involved quickly become prohibitive.

5.2.3 Pool-seq

Fortunately the GWAS requirement of genome wide polymorphism data from many individuals does not necessitate individual WGS; instead a pooled sequencing (pool-seq) approach can be taken. Rather than sequencing each individual to high depth of coverage, with each chromosome position represented multiple times, in pool-seq the DNA from multiple individuals is pooled and sequenced together, with much lower representation of individuals' chromosomes (Schlötterer *et al*., 2014). Sampling from a population usually involves measuring allele frequencies from a small part of it; therefore even high depth coverage (high accuracy of allele determination) of a few individuals can mean high allele sampling variance. Pool-seq allows many individuals to be sequenced for a fraction of the cost of traditional WGS which increases statistical power by producing a more accurate population allele frequency estimate (Figure 5.i1) (Futschik and Schlötterer, 2010).

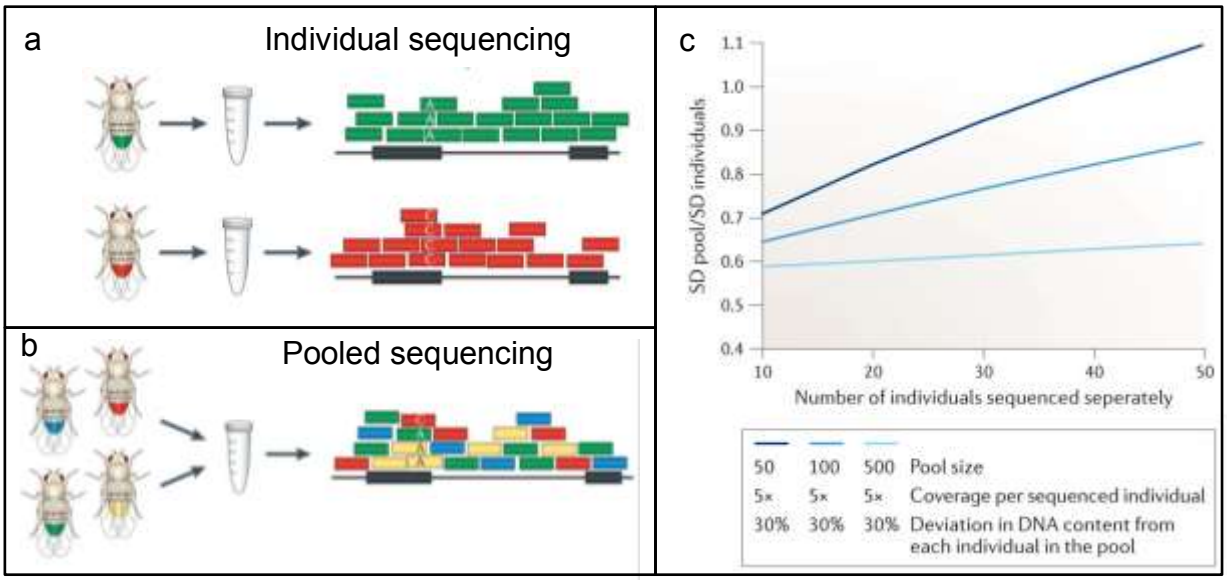


Figure 5.i1. Advantages of pool-seq. (**a**) WGS of two individuals from a fly population show two alleles present at locus in the population (A and C). (**b**) Pool-seq combines the DNA from four individuals and reveals two alleles present at the same locus (with more confidence) for around half the cost of the individual sequencing. Colours represent contributions from each individual to the sequenced reads (**c**) The accuracy of allele frequency estimation in pool-seq and individual WGS sequencing is compared using the standard deviation (SD) ratio of both methods, a value <1 means pool-seq is more accurate. A larger pool size increase the accuracy of estimation and individual sequencing accuracy only approaches and exceeds that of pool-seq when the number of individuals sequenced is close to that in the pool. Adapted from Schlötterer *et al*., 2014).

5.2.4 Pool-GWAS

By combining GWAS in wild populations with a pool-seq approach, accurate allelic comparisons between divergent phenotypes can cost-effectively establish associations (Schlötterer *et al*., 2014). The power of this approach was recently demonstrated in a study of pigmentation in European populations of *Drosophila melanogaster* (Bastide *et al*., 2013). Five pools of 100 *versus* 100 extreme phenotype individuals were analysed, revealing the association of SNPs in two genomic regions that had previously established to be regulatory for pigmentation genes (Bastide *et al*., 2013). The study provides an excellent proof of concept that a replicated pool-GWAS approach can establish convincing SNP level association with a trait in natural populations.

5.2.5 Aims

Recent advances in *A. arabiensis* genomic resources, including the availability of a chromosome arm construct reference genome (Sharakhov, Jiang and Hall, unpublished) and detoxification gene annotation (Neafsey *et al*., 2015) allow the exploitation of newly developed pool-GWAS techniques to elucidate insecticide resistance mechanisms in this important malaria vector. Replicates of resistance phenotyped pools of insects (resistant or susceptible) from a resistant Tanzanian population (Moshi) were compared with pools from a completely susceptible population (Tarime) (Kabula *et al*., 2012), with paired susceptible and resistant populations from the Zanzibar islands of Unguja and Pemba, respectively (Jones *et al*., 2013), offering further independent replication of associations. We aim to discover if, with a replicated within- and between-population approach, it is possible to isolate resistance candidate genes or SNPs in *A. arabiensis* for molecular validation and ultimately to develop resistance diagnostics for use in vector control.

5.3 Methods

5.3.1 Samples

Each pool was composed of 40 *A. arabiensis* females and was projected to provide a notional coverage of 1x per chromosome; a single *A. merus* pool included as an outgroup (Appendix 5.7.1). Mainland collections took place during the rice growing season in August-September 2012; Moshi alive and dead samples came from lower Mabogini (37° 21’ E 3° 24’ S), rice fields near lower Moshi on the southern slope of Mount Kilimanjaro, a region shown to have increasing resistance to pyrethroids (Matowo *et al*., 2014a). Mosquitoes were collected as larvae, raised to adults and females bioassayed in WHO tubes for one hour with 0.05% lambda cyhalothrin (WHO, 2013). Alive and dead mosquitoes were preserved over silica. In Tanzanian samples screened in Kabula *et al*., (2012), Moshi stood out as the most pyrethroid resistant population, though they were found to be completely DDT susceptible and only in one out of 642 mosquitoes assayed by Matowo *et al*. (2014a) was found to carry a *kdr* resistance mutation (*Vgsc-1014F*). Tarime collections took place in the village of Komaswa (34° 11’ E 1° 25’ S) about 410 km north west of Moshi, during July 2012. Mosquito larvae were collected, raised to adults and females bioassayed with a range of insecticides in WHO tubes for one hour (WHO, 2013), finding almost complete multi-insecticide susceptibility: permethrin (100% mortality), lambda cyhalothrin (97%), fenitrothion (100%), DDT (100%) and bendiocarb (100%). The 240 Tarime females sequenced in pools consisted of all the mosquitoes which died in pyrethroid bioassays (100 x permethrin and 97 x lambda cyhalothrin) with approximately equal numbers of dead from the other insecticides (Nyka, T. unpublished data – Insecticide resistance monitoring report 2012. NIMR Tanzania).

Samples from Zanzibar were collected during April and May 2012, with two collection sites on the island of Unguja (Mwera 39° 16’ E 6° 8’ S and Chuini 39° 13’ E 6° 5’ S) and two on Pemba (Tumbe 39° 47' E 4° 56' S and Mangwena 39° 43’ E 4° 58' S), these samples were collected as part of a study by Jones *et al*. (2013). Again, larvae were collected and raised to adults before females were bioassayed, using lambda cyhalothrin in WHO tubes for one hour (WHO, 2013). Results of bioassays were used to generate dose response curves and because resistance levels were found to differ drastically between island collections (10.7 – 13.3 fold) different insecticide concentrations were used accordingly (0.001%, 0.005%, 0.01%, 0.05%, 0.1% and 0.2%) (Jones *et al*., 2013). Mosquitoes from the lower end of the dose response curve in Unguja (the most susceptible) and from the highest end of the curve in Pemba (most resistant) were selected for sequencing (see Figure 2 in Jones *et al*., 2013). After DNA

extraction, all samples included in this study were identified as *A. arabiensis* using the SINE assay (Barnes *et al.,* 2005) and DNA concentrations were evaluated using Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher Scientific), taking the mean of two assays. Individuals were then pooled (N = 40 per pool) with equalised amounts of DNA.

5.3.2 Genomic alignment

The pools of *A. arabiensis* individuals were sequenced using Illumina technology (100bp paired end), by the Broad Institute. Each pool was split across 30 sequencing lanes to reduce lane biases and to ameliorate loses in case of technical failure, producing 30 .fasta files per pool. When these were downloaded, via ftp (file transfer protocol), they had been aligned to the *A. gambiae* reference genome (Holt *et al*., 2002). However as research into the genomics of the *Anopheles* species complex has revealed the benefits of aligning read to species specific reference genomes (Fontaine *et al*., 2015; Neafsey *et al*., 2015), these were re-aligned to two versions of the *A. arabiensis* reference genome. With many small contigs, the alignment to the *A. arabiensis* AraD1 reference genome available from VectorBase (Giraldo-Calderón *et al*., 2014), was not suited to sliding window analysis, as a single window could contain data from multiple contigs that bear no genomic proximity to one another potentially causing artefactual signals of divergence or masking true allelic differences between pools (see Discussion – Future Work). Fortunately, an *A. arabiensis* chromosomal reference (five large contigs, one for each chromosome arm) was developed (Sharakhov, Jiang and Hall, unpublished) using orthologue synteny with the *A. gambiae* AgamP3 reference genome (Holt *et al*., 2002), allowing read alignment to large stretches of contiguous sequence; we refer to this as the "AraChr" reference genome. The AraD1 alignment was used to investigate regions of the genome missing from the AraChr reference through lack of orthology with *A. gambiae.*

Before re-alignment, duplicate reads (possible PCR or sequencing errors) were removed from each file using MarkDuplicates from Picard-tools (http://picard.sourceforge. net), this was run before merging the lanes (into a single file per pool) as errors could be lane specific (command 1 – Appendix 5.7.2). Data from individual lanes were then merged using Picard-tools MergeSamFiles (Command 2 – Appendix 5.7.2), before being converted into .fastq format with Picard-tools SamToFastq (Command 3 – Appendix 5.7.2). These reads were then re-aligned to the AraChr reference genome with BWA sampler (Li and Durbin, 2009) (Command 4 – Appendix 5.7.2). Low mapping quality reads (MAQ<20), those with poor or

ambiguous alignment to the reference (Li, Ruan and Durbin, 2008), were removed and the files converted from the human readable .sam format into the much smaller sized binary .bam format using Samtools view (Command 5, Appendix 5.7.2) (Li *et al*., 2009). The mean depth of sequencing coverage was calculated for each pool using the genomeCoverage function from BEDtools (Command 6, Appendix 5.7.2) (Quinlan, 2014).

5.3.3 Pool-seq analysis

With WGS of individuals, genotypes are calculated for each specimen, however with a pooled sequencing approach, the frequencies of alleles at each locus are instead used for population genomic analyses; the software package PoPoolation2 provides a framework and amenable data format for this type of work (Kofler, Pandey and Schlötterer, 2011). Data from pools are first collated and reformatted as .mpileup using Samtools mpileup (Command 7 – Appendix 5.7.2) (Li *et al*., 2009), before being converted into a .sync file, using the mpileup2sync.jar within PoPoolation2 (Command 8 – Appendix 5.7.2) (Kofler, Pandey and Schlötterer, 2011). The Java version of this function was used as it computes much faster than the Perl version, allowing multithreading. At this step, an individual base quality filter, removing bases with quality < 20, is applied to remove low quality bases.

Sync files were subsampled to normalise the results across pool sequencing depths. With 40 diploid individuals in each pool, 80 alleles were randomly chosen, without replacement, for each position using (Command 9 - Appendix 5.7.2), except in the case of samples from Zanzibar, Unguja pool sequencing suffered some technical failure andhad approximately half the depth of coverage so all Zanzibar pools were subsampled at 40 alleles to normalise allele frequencies (Appendix 5.7.1). This function also acts as a minimum depth so positions where any also pool has lower than target depth are ignored as are positions that fall in the top 2% of highest depth, as these may be alignment issues, for example misalignment of repeated genomic regions which could inflate polymorphism at a locus. These data preparation steps follow the pool-seq best practices from Schlötterer *et al*. (2014). Finally, a minor allele frequency cut off filter of 5% was applied in order to remove loci uninformative about differences between pools and treatments, using a custom Perl script (https://github.com/cclarkson/thesis_ chapter_5/blob/master/sync_MAFfer.pl).

5.3.4 Allele frequency probabilities

The sync files for mainland Tanzania *A. arabiensis* pools were randomly assigned to two equally sized replicates for each treatment (*e.g.* Moshi alive 1, Moshi alive 2) using a custom Perl script (https://github.com/cclarkson/thesis_chapter_5/blob/master/random_pool_picker.pl). Zanzibar, with just four pools (mainland has 22) had just one pool per replicate and, because there was only one pool of *A. merus*, all comparisons with it were calculated with no replicates *i.e.* All Moshi dead pools *versus A. merus* (Figure 5.1). To calculate the p-values of pairwise allele frequency differences between groups of pools, chi-square tests were conducted at each variant passing the filtering (see **Pool-seq analysis**)using a custom Perl script (https://github.com/ cclarkson / thesis_chapter_5/ blob/master/chi2_*.pl).

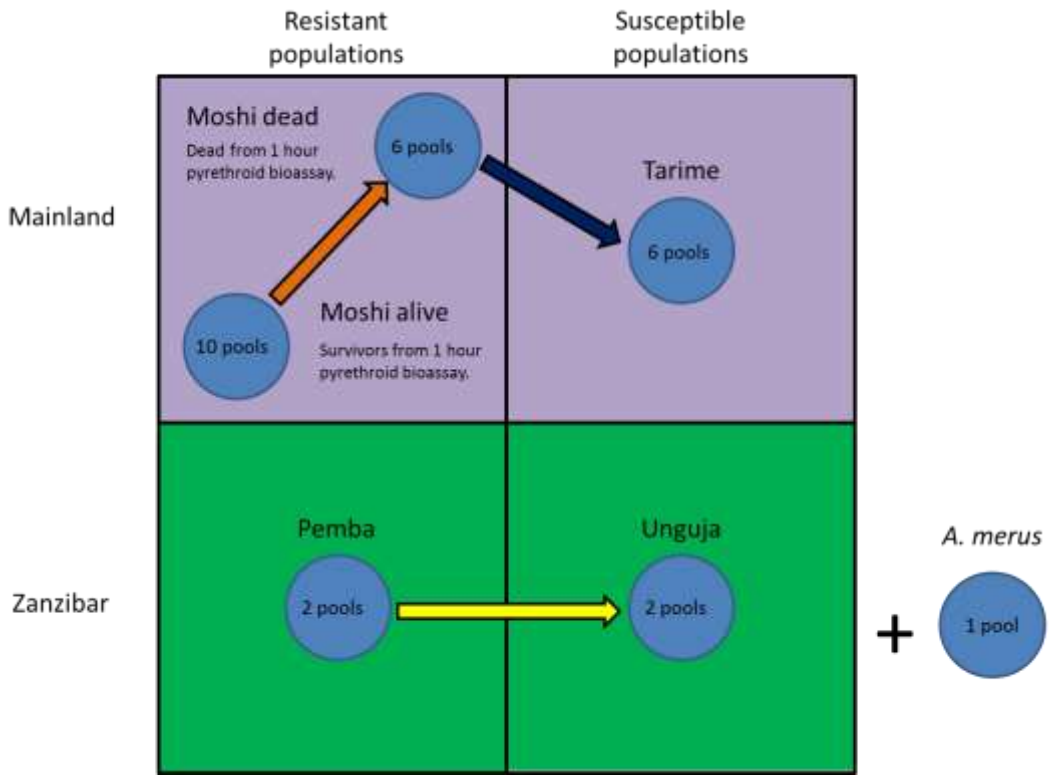


Figure 5.1. Experimental design. Blue circles show the groupings of pools, with resistance status of population, location and number of pools shown for each group. Arrows show the direction of expectation in reduction of candidate association to resistant phenotype.

For data aligned to the AraChr reference a windowed approach was adopted to account for linkage between variants and to reduce the number of data points for ease of visualisation. A non-overlapping window of 1000 SNPs was slid along each chromosome arm with

probabilities and summary statistics calculated for each window (mean, variance, geometric mean, median and minimum) using another custom Perl script (https://github.com/ cclarkson/thesis_chapter_5/blob/master/windowed_ probabilities.pl). A SNP based, rather than nucleotide based, window was used to ensure mean probabilities were not just being driven by a small number of variants; however this was at the cost of a uniform genomic size of window.

5.3.5 Candidate filtering

To define insecticide resistant candidate windows and SNPs, a number of filtering rules were applied to the chi-squared probabilities. Probabilities were ranked then converted to percentiles bounded between one and zero, by dividing by the number of variants or windows (depending on which was being analysed). A threshold was then applied to take the best hits in resistance susceptible comparisons. From the mainland comparisons a 0.05 percentile was applied (intra-population -Moshi alive *versus* dead and inter-population - Moshi dead *versus* Tarime) and from Zanzibar (Unguja *versus* Pemba) the 0.01 percentile was used (Figure 5.2). Moshi dead was used in the Tarime comparison due to independence of expectation in a window/variant's association to phenotype, allowing the directionality to candidate association expectation: Tarime < Moshi dead < Moshi alive, to be evaluated (Figure 5.1) (Edi *et al*., 2014). To be considered as mainland candidates, windows or SNPs had to have lower probabilities in both replicates of resistance *versus* susceptible in both intra (Moshi *versus* Moshi) and inter-population (Moshi *versus* Tarime) comparisons than in the replicates of resistant *versus* resistant and susceptible *versus* susceptible in both intra and inter-population analyses. A higher threshold was therefore set in Zanzibar (0.01 percentile), to compensate for the lack of independent intra-population comparison, as here a resistance candidate just had to have lower probabilities in both replicates of Unguja (susceptible) *versus* Pemba (resistant) than both Unguja *versus* Unguja and Pemba *versus* Pemba. Regions concordant (overlapping) across both mainland and Zanzibar analyses were considered the strongest hits. The statistical software R was use to plot all data in the study (R Development Core Team, 2014).

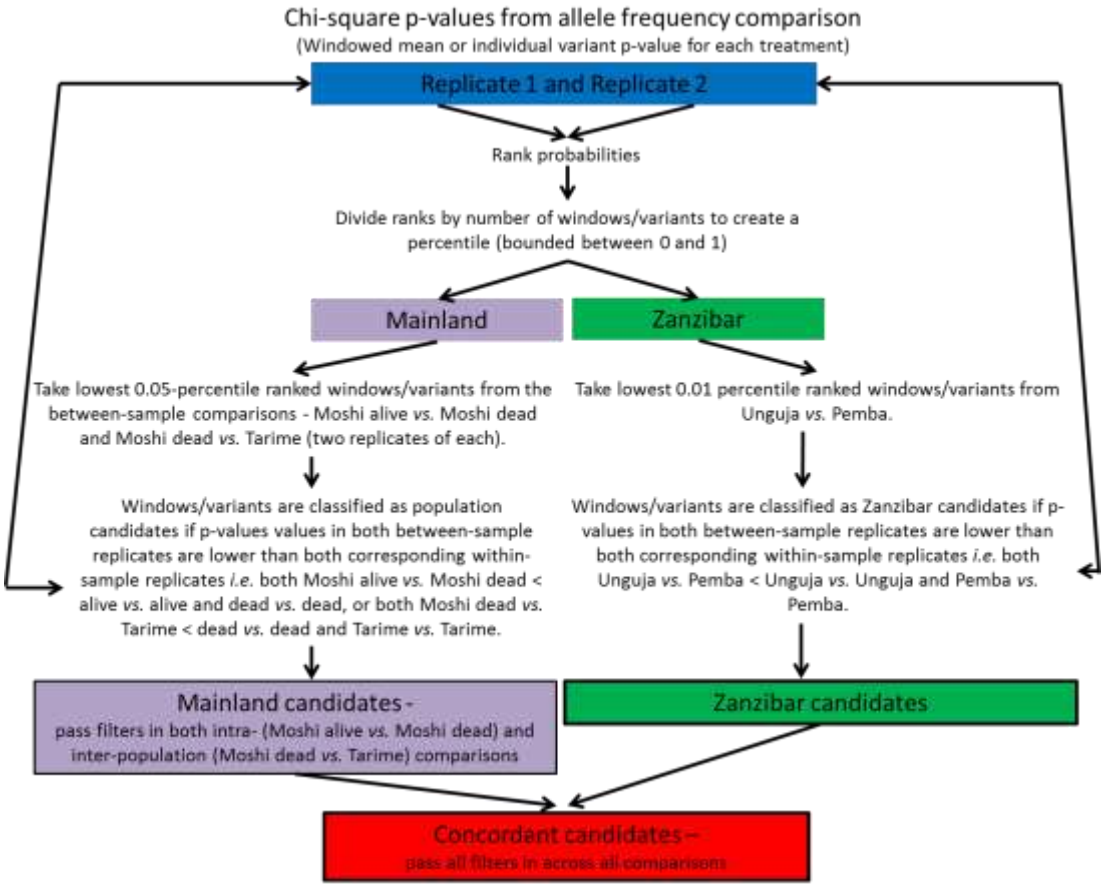


Figure 5.2 Flow diagram of the candidate window and variant filtering process.

5.3.6 Identification of non-synonymous mutations

Recent research into the genomics of the *Anopheles* complex involved genes from families known to be linked to insecticide resistance, including cytochrome P450 and glutathione S-transferase genes, being manually annotated for *A. arabiensis* (Neafsey *et al*., 2015). Effects of candidate SNPs that fell within the coding regions of these genes were analysed for their possible functional effect and to identify potential regulatory SNPs. To locate the position of these genes' exons in the AraChr reference genome, the exon sequences were extracted from the AraD1 reference genome using the AraD1.3 gene set from VectorBase (Giraldo-Calderón *et al*., 2014) and BLASTed (Camacho *et al*., 2009) against a database made from the AraChr reference; positions were then taken as the top hit (100% identity) for each exon. For variants identified as resistance associated candidates and located within annotated exons, their positions were converted back to AraD1 (calculated as AraChr position minus 18354335bp) and run through Variant Effect Predictor on VectorBase using the AraD1 reference to give the variant functional effects (McLaren *et al*., 2010; Giraldo-Calderón *et al*., 2014).

5.3.7 Copy number variants

Genomic duplications or copy number variants (CNVs) may inflate apparent allelic variation at loci through incorrect alignment to reference genomes due to the similarity of recently duplicated regions. To test for the incidence of duplication driving divergence we measured the depth of coverage at each variant using the allele frequencies taken from the raw, unfiltered .sync file and a custom python script (https://github.com/cclarkson/thesis_chapter_5/blob/master/window_group_depth.py) which generated a 1000 variant stepping window mean depth for each grouping of mainland pools (MA1, MA2, MD1, MD2, T1 and T2).

5.3.8 Investigation of structural features using *A. merus*

On the 2R chromosome arm, *A. arabiensis* sports a number of polymorphic inversions (Coluzzi *et al*., 2002). The reduction in recombination between alternate inversion types may generate divergence in the absence of selection (Marsden *et al*., 2014). *A. merus*, however, is fixed for inversions in this genomic region, so a pool of 40 females from Zanzibar was also sequenced and aligned to the *A. arabiensis* AraChr reference to allow comparison to the other pools. Both because the *A. merus* pool had the lowest mean sequencing coverage (Appendix 5.7.1) and due to divergence from *A. arabiensis*, alignment to a non-specific reference is sub-optimal in term of number of reads mapping (Fontaine *et al*., 2015; Neafsey *et al*., 2015). Therefore a second data set was produced for this analysis. Popoolation2 was used in the same way as above but the inclusion of the *A. merus* pool meant that when subsampling to 80 alleles, a smaller, lower resolution dataset was produced because any pool with less than 80 alleles at a variant causes that position to be lost from the data set (Kofler, Pandey and Schlötterer, 2011). Though this reduced the number of variants by 76.53%, 107242 variants were left on the 2R chromosome arm for comparison of *A. merus* with *A. arabiensis* pools.

5.3.9 Specific candidate investigation - *Cyp4g16*

Recent molecular investigations into the genetic causes of pyrethroid resistance in Tanzania, found that upregulation of a cytochrome P450 gene, *Cyp4g16*, was linked with resistance to the insecticide (Jones *et al*., 2013; Matowo *et al*., 2014b). The AraD1 reference genome contig carrying this gene (KB704462) did not contain enough genes to be included in the AraChr reference, and so was analysed using the AraD1-aligned dataset. As the contig was

only 22284bp, allele frequency probabilities were analysed for variants individually rather than using a windowed approach.

5.4 Results

5.4.1 Genome wide association

The experimental design of this GWAS allows not only replication between and within mainland Tanzanian insecticide resistance and susceptible populations of *A. arabiensis* but also between geographically-isolated island populations from Zanzibar. If present, resistance mechanisms shared across sampling sites, with common or convergent origins, would be revealed as concordant candidates. Three pairwise population comparisons were explored here in depth. A genome scan of the allele frequency differences between replicates of Moshi alive (survivors of WHO bioassay) and Moshi dead (those that didn't survive) revealed low probabilities across much of the chromosome arms as perhaps expected with both coming from the same sampling location. However a defined peak of allelic difference is clearly visible on the 2R chromosome arm between both replicates of alive *versus* dead pools (Figure 5.3a). Filtering of the Moshi alive *versus* dead geometric mean probabilities produced 37 candidate windows at the 0.05 percentile, across all chromosome arms (Table 5.1).

Table 5.1. Insecticide resistance candidate windows. Table shows the total number of 1000 variant windows on each chromosome arm and in total, for mainland comparisons (Moshi alive *versus* Moshi dead and Moshi dead *versus* Tarime) and Zanzibar (Unguja *versus* Pemba). Intra refers to within-population comparisons and inter, between populations, while the numeric value shows the percentile threshold for inclusion in filtering.

	Mainland				**Zanzibar**		**Combined**
chr	**windows**	**intra 0.05**	**inter 0.05**	**intra + inter**	**Windows**	**inter 0.01**	**mainland + Zanzibar**
2R	456	15	20	4	1098	9	1
2L	361	7	15	0	850	4	0
3R	385	9	16	0	866	6	0
3L	302	5	11	1	671	6	0
X	103	1	4	0	217	2	0
total	1607	37	66	5	3702	27	1

Peaks of allele frequency probabilities were more striking between Moshi dead, which, though they did not survive a 60 minute WHO lambda cyhalothrin insecticide bioassay were sampled from a population with high pyrethroid insecticide resistance (Kabula *et al*., 2012), and Tarime, a geographically distinct population susceptible to all insecticides (Nyka, T. unpublished data – Insecticide resistance monitoring report 2012. NIMR Tanzania). Again, the largest peak can be seen on 2R, concordant with that found when Moshi alive and dead were compared, though other peaks are evident across the genome (Figure 5.3b). Filtering of these data, at the 0.05 percentile, highlighted 66 candidate windows; of those on the 2R chromosome arm four candidate windows were both contiguous and concordant with four candidate windows from the Moshi alive *versus* dead analysis (Table 1), spanning the region 2R: 29163023-29865540 (AraChr reference). One other window was concordant across Moshi and Tarime on the 3L arm.

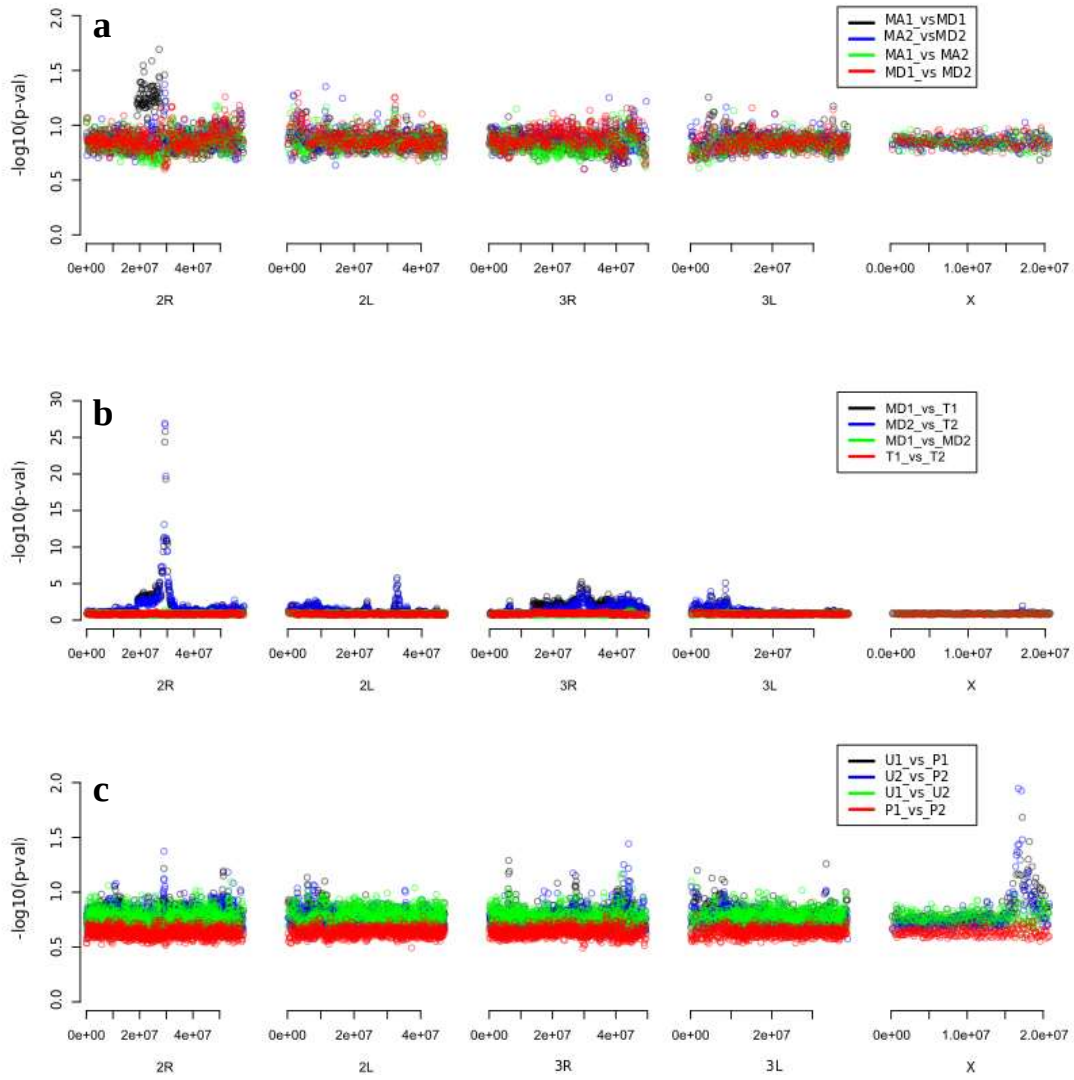


Figure 5.3. Resistant versus susceptible samples pairwise mean p-values across the *A. arabiensis* genome. Each open circle represents the relationship between the geometric mean –log10(p-value) for a 1000 SNP stepping window (chi^2 test on allele frequencies between groups of pools) plotted as the window mid-point nucleotide position along the chromosome arm. (**a**) Moshi alive *versus* Moshi dead. (**b**) Moshi dead *versus* Tarime. (**c**) Unguja *versus* Pemba. Boxed legend for each plot shows population point colour: MA = Moshi Alive, MD = Moshi dead, T = Tarime, U = Unguja, P = Pemba. Number following population code represents replicate (1 or 2). Note - y-axis scale differs between plots.

Pairwise comparisons of the Zanzibar island populations (Unguja – susceptible, Pemba, resistant) revealed a probability landscape with similar amplitude to the within population Moshi alive *versus* dead, suggesting little divergence despite island isolation and a stronger

phenotype selection (Jones *et al*., 2013). Though these are difficult to compare directly as the pool size is smaller in Zanzibar samples, which will impact p-values. The Zanzibar susceptible *versus* resistant replicates (U1_vs_P1, U2_vs_P2) showed peaks on every chromosome arm, however in contrast with both mainland comparisons, the strongest signal was apparent on the X chromosome (Figure 5.4c). Mainland replicates, with an intra-population (Moshi alive *versus* Moshi dead) and an inter-population (Moshi dead *versus* Tarime) analysis, allowed a conservative two-step candidate filtering (concordance required in both intra and inter) As Zanzibar sampling provided only inter-population filtering, we limited candidate windows to those with a lower percentile threshold of 0.01, generating 27 insecticide resistance candidate windows across all chromosome arms. One window on the 2R chromosome arm overlapped with a candidate window from the mainland analysis and, as the only region concordant across all three comparisons, was the strongest replicated insecticide resistance candidate region for further investigation: 2R:29161471-29386646 (AraChr reference).

5.4.2 Inversions

When comparing samples from Cameroon and central Tanzania, a previous study evaluating the prospects for GWAS in *A. arabiensis* also found genetic differentiation in a region of the 2R chromosome arm near to the candidate window we identified (Marsden *et al*., 2014). The authors suggested that reductions in recombination caused by the 2Rb and 2Rc inversions, polymorphic in *A. arabiensis* (Coluzzi *et al*., 2002), may drive this divergence (Marsden *et al*., 2014). It is difficult to compare the positions of these accurately, because the previous study aligned sequencing reads to the *A. gambiae* reference genome (Holt *et al*., 2002), rather than *A. arabiensis*. However, evidence of the 2Rb inversion (Coluzzi *et al*., 2002; Marsden *et al*., 2014) is clearly suggested in both Moshi dead *versus* Moshi alive (Figure 5.4a – MA1_vs_MD1) and Moshi dead *versus* Tarime comparisons (Figure 5.4b - MD1_v_T1 and MD2_v_T2). This finding suggests the 2Rb inversion is polymorphic in Moshi but fixed in Tarime. The striking peaks in probability appear lie within a region telomeric to this, covered by the 2Rc inversion (Figure 5.4) (Coluzzi *et al*., 2002), however in all but Moshi dead *versus* Tarime (Figure 5.4b) the peak appears to only cover part of the inversion and the divergence is found only in resistant versus susceptible replicates in each comparison (Figure 5.4a, c).

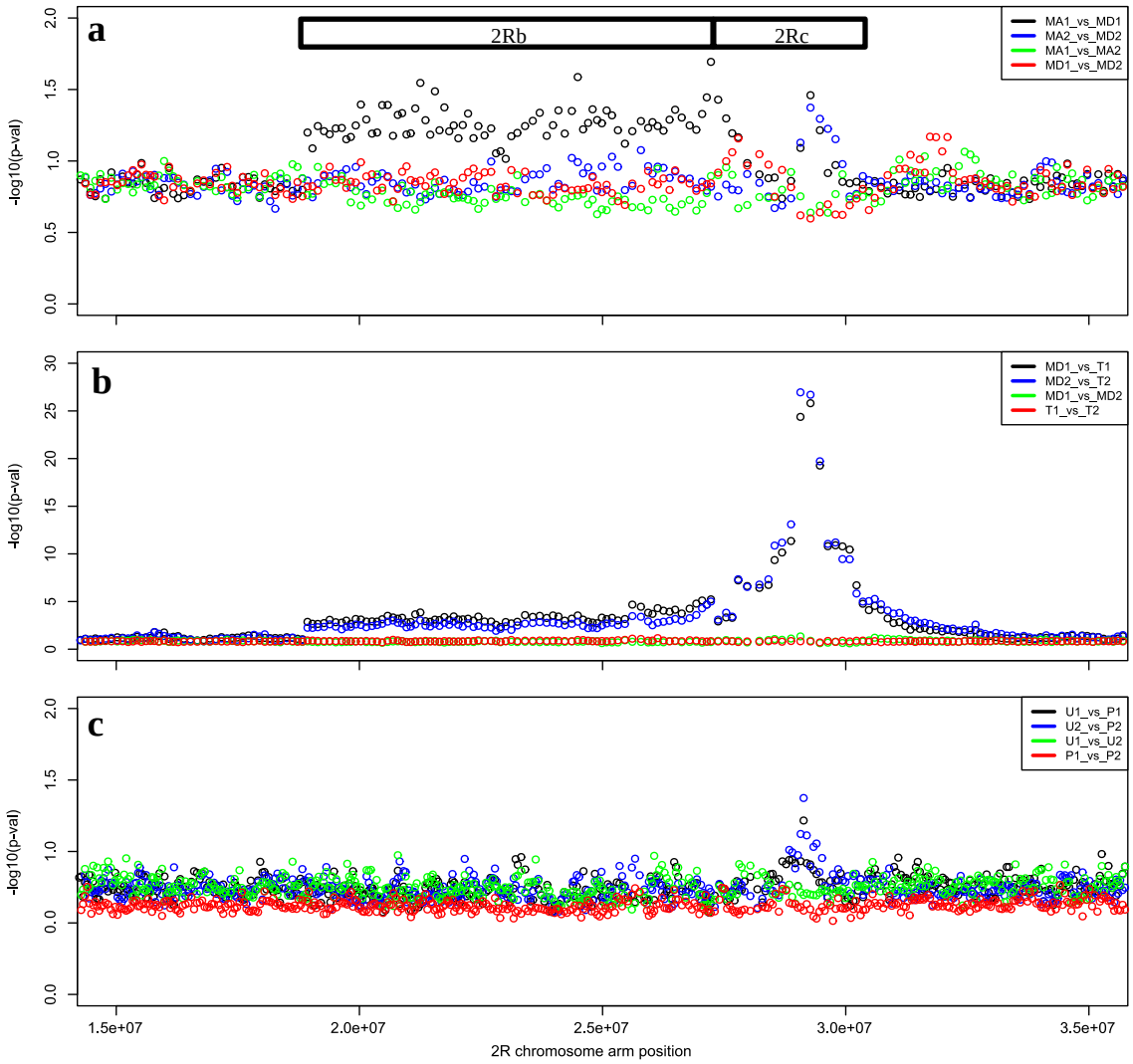


Figure 5.4. 2R "peak" region pairwise mean p-values across the *A. arabiensis* genome. Each open circle represents the relationship between the geometric mean –log10(p-value) for a 1000 SNP stepping window (chi^2 test on allele frequencies between groups of pools) and its mid-point nucleotide position along the 2R chromosome arm. (**a**) Moshi alive *versus* Moshi dead. (**b**) Moshi dead *versus* Tarime. (**c**) Unguja *versus* Pemba. Boxed legend for each plot shows population point colour: MA = Moshi Alive, MD = Moshi dead, T = Tarime, U = Unguja, P = Pemba. Number following population code represents replicate (1 or 2). Black outlined bar in (**a**) denotes approximated positions of 2R polymorphic inversions in *A. arabiensis*. Note - y-axis scale differs between plots.

5.4.3 2R candidate region genes and SNPs

To ensure any regulatory regions were not missed and to compensate for window edge effects that may have affected windowed means, the 2R candidate region explored was

extended the width of the region (225,175bp) in both centromeric and telomeric direction. Using a smaller stepping window allowed a higher resolution view of this 2R extended insecticide resistance candidate region. Visualisation of 100 variant windows revealed that the 2R peak of allelic divergence in all resistant *versus* susceptible comparisons was found to overlap a cluster of cytochrome P450 genes from the CYP6 family (Figure 5.5), some of which are known to be involved in detoxification of pyrethroids in *Anopheles* (Müller *et al*., 2008; Witzig *et al*., 2013) and other insects (McCart and ffrench-Constant, 2008; Chiu *et al*., 2008). Within the candidate region itself, four genes from the family are found: *Cyp6p1, Cyp6p2, Cyp6p4* and *Cyp6ad1*. A full list of the 22 genes found in the candidate region (49 including extended candidate region) can be found in Appendix 5.7.3. Unfortunately a higher resolution view using a smaller window did not localise any increased probability on a specific region or gene in this CYP450 cluster (Figure 5.5 – inset). Within the extended candidate region 141 mainland and 508 Zanzibar individual variants were present. After filtering the variants in the same fashion as the windows, none were found concordant between mainland comparisons at the 0.05 percentile, nor between mainland 0.05 percentile and Zanzibar 0.01 percentile (Table 5.2). However, it was possible to investigate the effects of variants that were significant within comparisons on the genes in the CYP450 cluster due to the recent update of the AraD1.3 gene set with the manual annotations of known detoxification genes (from the 16 genomes project – Neafsey *et* al., 2015) on VectorBase (Giraldo-Calderón *et al*., 2014).

The effect of SNPs that were identified as candidates within comparisons were investigated using the software VEP (McLaren *et al*., 2010). Of those falling in exonic regions only three caused amino acid substitutions. Two mutations, one a candidate for resistance on the mainland (Moshi dead *versus* Tarime) and one for Zanzibar, hit *Cyp6p2*, with the mainland candidate causing the stop codon to be lost, possibly causing the protein to be non-functional. However, the population with the highest loss of function allele frequency on the mainland was resistant (Moshi dead), suggesting that if the gene was involved in resistance then it was not carrying a functional gene that conferred resistance, this result, therefore, appears unlikely to be related to a resistance phenotype. The last non-synonymous mutation, a stop-lost, affects *Cyp6p15p*, a candidate between Moshi dead *versus* Tarime pools (though falling in the extended region). Further molecular investigation may be warranted as allele frequencies indicate Tarime pools (susceptible) appeared to have a dearth of functional *Cyp6p15p* genes, 424:56 (stop-lost:wild type allele) compared to 39:441 in Moshi dead.

Table 5.2. Variant Effect Predictor results for variants within the extended 2R candidate region which are associated with resistance phenotype. The 'concordance' column reports which pairwise resistant *vs.* susceptible comparison the association lay. * signifies a variant that is in the 0.05 percentile for inter-population comparison (Moshi alive *vs.* Moshi dead) and passes the filters for intra-pop (Moshi dead *vs.* Tarime) but did not fall in the intra-population 0.05 percentile. The extended region covers the concordant candidate region plus ~225bp up and downstream. The carboxylesterase alpha esterase was not manually annotated but named via orthology with *A. gambiae* on VectorBase. All genes are on the negative strand.

position	allele	Effect	Gene	VectorBase	Concordance
29143211	T/C	non-coding			Zanzibar
29152653	T/C	non-coding			Zanzibar
29153262	G/T	synonymous	carboxylesterase alpha esterase	AARA015786-RA	Intra
29153343	A/G	synonymous	carboxylesterase alpha esterase	AARA015786-RA	Inter
29154076	C/T	non-coding			Inter
29154628	A/G	stop_lost	CYP6P15P	AARA015785-RA	Inter
29160141	A/G	non-coding			Zanzibar
29160638	C/G	non-coding			Intra
29161351	A/T	non-coding /splice variant	CYP6P5	AARA015788-RA	Inter
29164251	G/A	synonymous	CYP6P4	AARA015789-RA	Intra
29167811	G/A	non-coding			inter*
29168615	C/T	Synonymous	CYP6P2	AARA015791-RA	Zanzibar
29168846	A/G	stop-lost	CYP6P2	AARA015791-RA	inter
29169229	G/A	Missense	CYP6P2	AARA015791-RA	Zanzibar
29176579	C/A	non-coding			intra
29176626	T/C	non-coding			intra

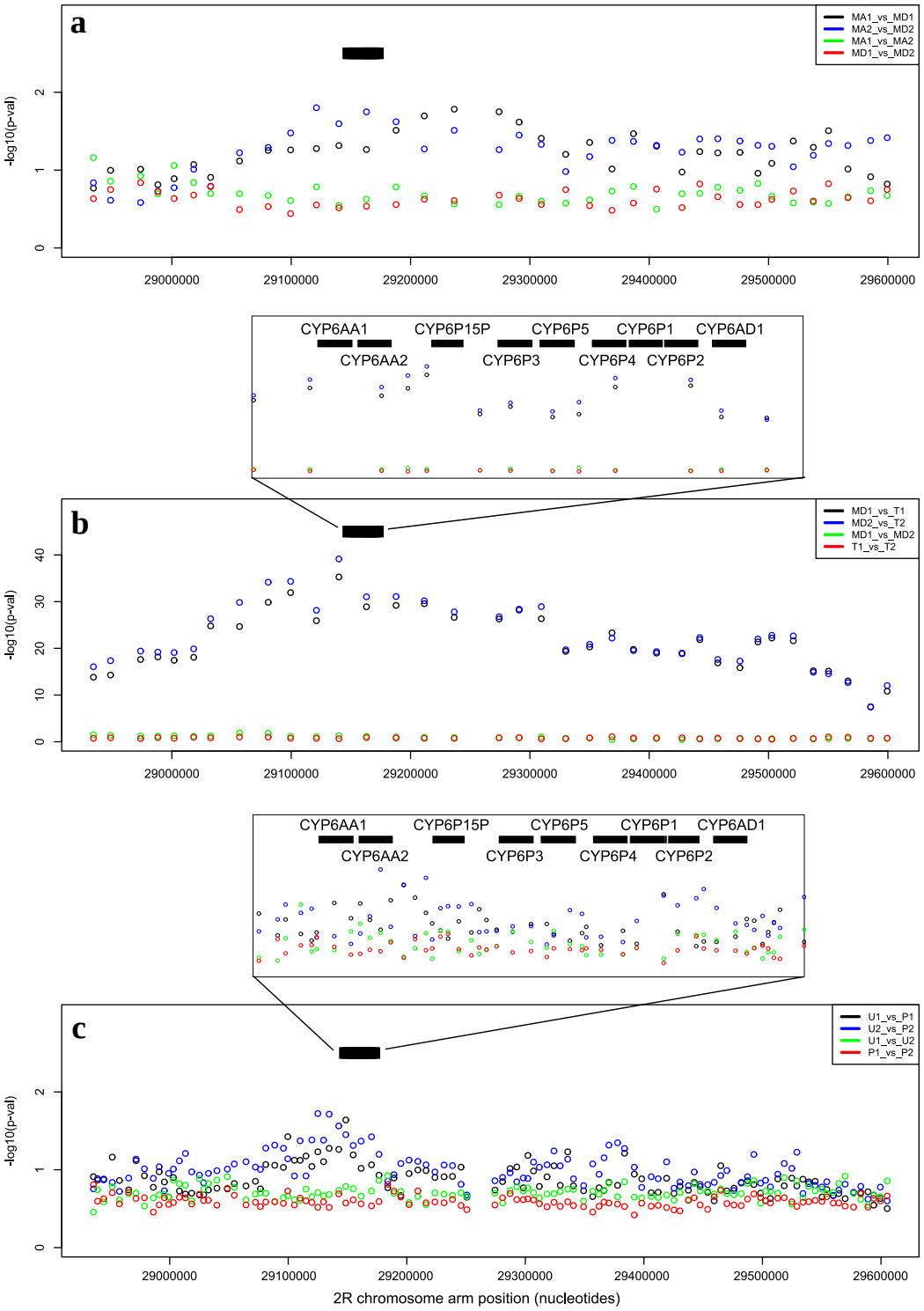


Figure 5.5. 2R 'peak' region and cytochrome p450 cluster pairwise p-values. In main figure each open circle represents the relationship between the geometric mean –log10(p value) of a 100 variant stepping window (chi^2 test on allele frequencies between groups of pools) and window mid-point along the CYP450 cluster overlapping the candidate region on the 2R chromosome arm. The black bar above points shows the position of a cytochrome P450 gene cluster of 9 genes. Inset figures show the geometric mean –log10(p-value) for 10

variant stepping windows over the cytochrome P450 cluster, here black bars show genes. (**a**) Moshi alive *versus* Moshi dead. (**b**) Moshi dead *versus* Tarime. (**c**) Unguja *versus* Pemba. Boxed legend for each plot shows population point colour: MA = Moshi Alive, MD = Moshi dead, T = Tarime, U = Unguja, P = Pemba. Number following population code represents replicate (1 or 2). Note - y-axis scale differs between plots.

5.4.4 Copy number variants

One possible explanation for the allelic divergence peaks between resistant and susceptible populations (Figure 5.3) is that directional selection on candidates has driven the evolution of duplicated regions of the genome in *A. arabiensis.* Duplication of functionally constrained genes allows rapid adaptation without the potential fitness costs associated with point mutations (Ohta, 1989), for example the loss in fitness associated with carrying the knock down insecticide resistance (*kdr*) mutations (Foster *et* al., 2003; Brito *et al*., 2013) may be linked with duplications found in the voltage gated sodium channel gene (VGSC) in *Culex quinquefasciatus* (Xu *et al*., 2011). In *A. gambiae* an increase in the number of copies of the *Ace-1* gene coding for synaptic acetylcholinesterase, a target for some insecticides utilized in vector control, has been found (Djogbénou *et al*., 2008b; Weetman *et al*., 2015), and this copy number increase has been linked to increased insecticide resistance (Edi *et al*., 2014). To investigate this, the depth of sequencing coverage was calculated across the 2R chromosome arm for each mainland population (Moshi alive, Moshi dead, Tarime).

Due to the similarity of recently duplicated genomic regions to their progenitor, sequencing reads from CNVs may align to the same place on the reference genome and therefore increase the apparent depth of a region (Yoon *et al.*, 2009). This could alter divergence by certain populations having more allelic diversity at a duplicated locus than populations without the CNV. A similar effect is found with transposable elements which are often highly duplicated (Li *et al*., 2005). Figure 5.6a demonstrates a number of these higher coverage regions are found across the 2R chromosome arm, possibly driven by CNVs. Perhaps the most distinct of these high coverage peaks, found in both the Moshi and Tarime populations (Figure 5.6a – black arrow), falls proximate to the insecticide resistance candidate region on 2R generated by the GWAS around 30Mb from the centromere (Figure 5.3). However, closer inspection reveals that the potential duplication (Figure 5.6b – black arrows) is ~500kb distant, in the centromeric direction, to the peak in allele frequency and to the cluster of cytochrome P450 genes (Figure 5.6b black rectangle). A higher resolution

investigation also shows that this 'peak' in depth is actually a region with highly variable depth, with windows revealing very high and very low values (Figure 5.6b). These findings suggest that the allele frequency divergence peak found between insecticide resistant and susceptible populations on the 2R chromosome arm is not being driven by CNVs.

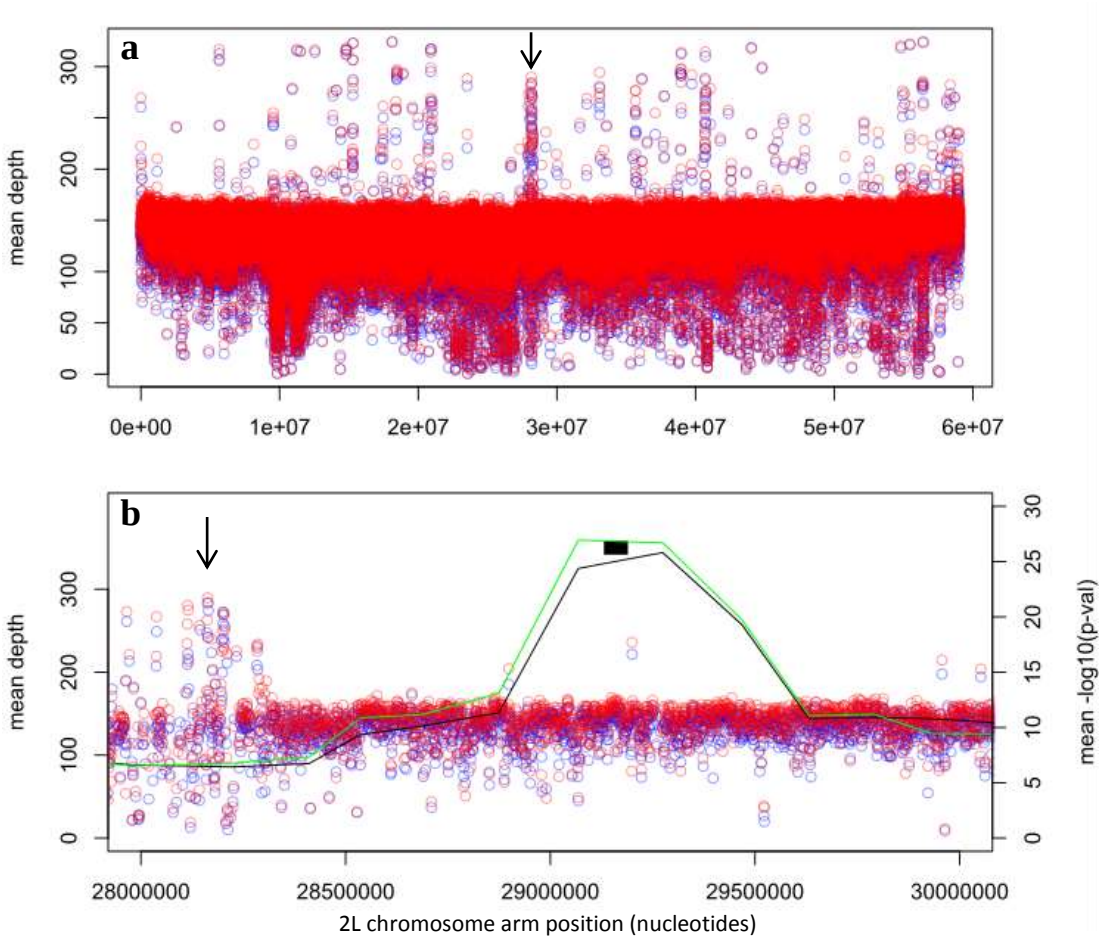


Figure 5.6. Moshi dead and Tarime mean depth of sequencing coverage over 2R chromosome arm. Open circles represent the mean depth of sequencing coverage across Moshi dead (blue) and Tarime (red) pools, in 1000 variant stepping windows (**a**) Figure shows the relationship between mean depth (pre-subsampling) and window mid-point on the 2R chromosome arm. (**b**) A close up of the approximate 2R chromosomal region containing both the peak in coverage at ~28Mb (from **a**), with mean depth (y axis), and the mean –log10 (p-value) in 1000 variant stepping windows of chi^2 tests of allele frequency between pool replicates (z axis): Moshi dead replicate 1 *versus* Tarime replicate 1 (black line) and Moshi dead replicate 2 *versus* Tarime replicate 2 (green line). Black rectangle shows position of the 2R cytochrome P450 cluster.

5.4.5 Investigation of structural features using *A. merus*

The candidate region explored on the 2R arm may lie within the 2Rc inversion. Therefore, if inversion orientation was driving insecticide resistant phenotypes in resistant populations, a difference in allelic difference topography might be expected when comparing resistant *versus A. merus* and susceptible *versus A. merus* as *A. merus* is fixed for inversions on 2R (Coluzzi *et al*., 2002). Unfortunately *A. merus* was so divergent from *A. arabiensis* that alignment of reads to the *A. arabiensis* AraChr reference saw a drop of over 75% in the number of variants with allelic information (after filtering). This reduced density of variants, compared to within *A. arabiensis* comparisons and left gaps in the data rendering it difficult to discern differences between resistant populations *versus A. merus* and the susceptible comparison (Appendix 5.7.4).

5.4.6 Specific candidate investigation - *Cyp4g16*

Cyp4g16, a gene previously linked with *A. arabiensis* pyrethroid resistance in both mainland (Matowo *et al*., 2014b) and Zanzibar populations, (Jones *et al*., 2013), did not appear in the AraChr alignment so was investigated separately. As expression variation of the gene had previously been associated with resistance, a non-coding variant was hypothesised; however, pairwise allele frequency analysis between susceptible and resistant replicates found no single variant strongly pyrethroid resistance associated (Figure 5.7). A lack of concordance across comparisons suggests that if the expression variant is *cis* regulated (rather than *trans* which we could not detect), the regulatory variant may be different between populations and/or perhaps falls with the coverage gap in the gene (Figure 5.7).

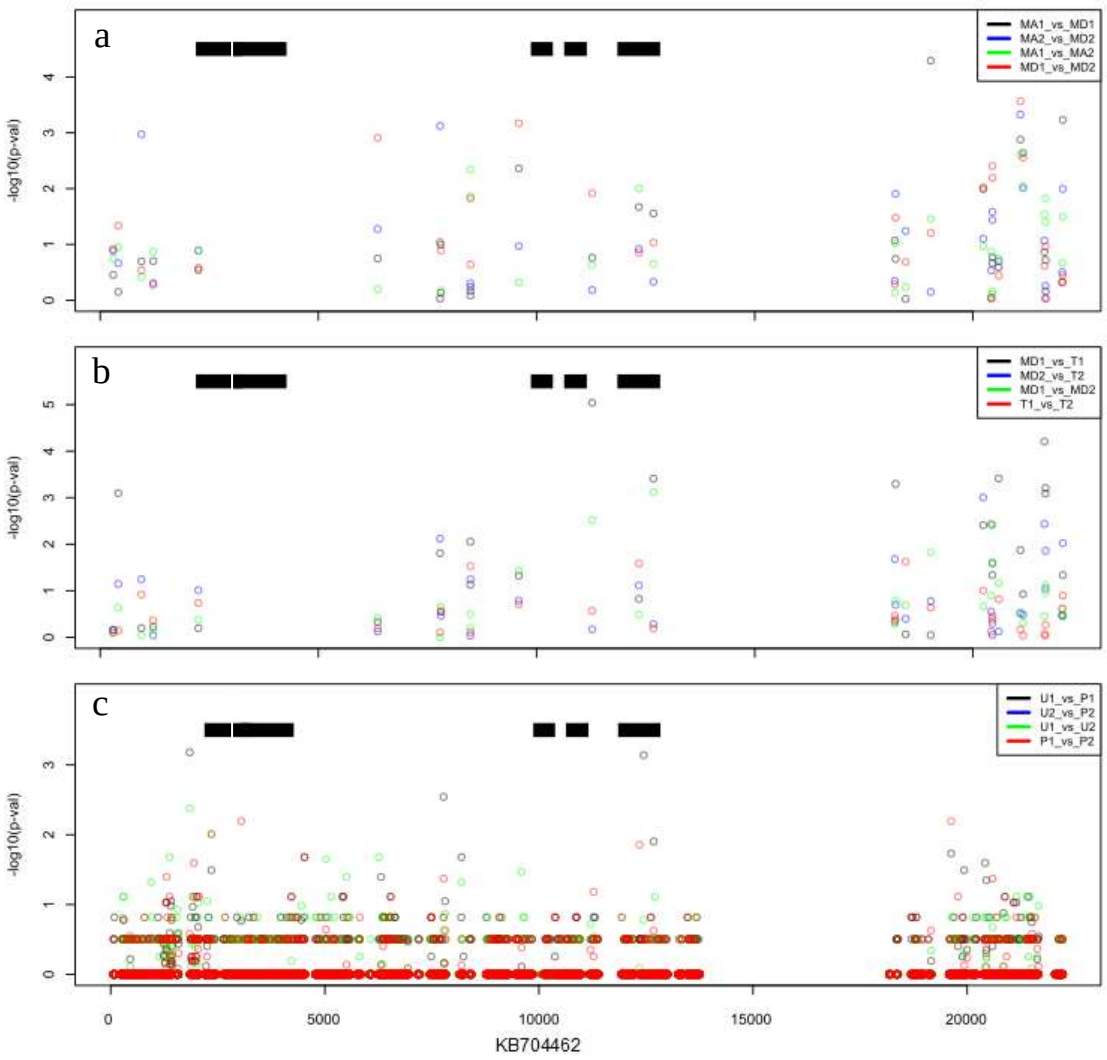


Figure 5.7. Pairwise p-values for *A. arabiensis* comparisons across contig KB704462. (**a**) Moshi alive *versus* Moshi dead. (**b**) Moshi dead *versus* Tarime. (**c**) Unguja *versus* Pemba. Black bars show location of *Cyp4g16* exons.

5.5 Discussion

5.5.1 Pool-GWAS

A pool-seq approach enabled genome wide allele frequencies from over 1000 mosquitoes to be analysed for association with a pyrethroid insecticide resistance phenotype, for a fraction of the cost of using WGS individual genomes. This large sample generated 23 independent pools of 40 individuals, allowing replicates of insecticide resistant and susceptible *A. arabiensis* over three pairwise comparisons: both intra-population, between survivors and non-survivors of a lambda cyhalothrin bioassay in Moshi (Tanzania), and inter-population, between Moshi and a susceptible population in Tarime (Tanzania) and between resistant and susceptible island populations from Zanzibar. The experimental design enabled levels of confidence in candidate regions through concordance across replicates and comparisons. Within comparisons, multiple insecticide resistant candidates were evident across the genome; however, the strongest candidate was a ~225kb region on the 2R chromosome arm concordantly associated with pyrethroid resistance across all comparisons.

The patterns of divergence across the 2R chromosome arm suggest that this candidate lies within the 2Rc inversion, polymorphic in *A. arabiensis*, possibly close to the 2Rb/2Rc inversion breakpoint (Coluzzi *et al*., 2002). A recent WGS GWAS, investigating the prospects for association mapping in *A. arabiensis* found differentiation between populations in this region using the fixation index F_{ST}, the authors suggest it may be driven by genetic drift via reduced recombination between standard and inverted arrangements of inversion (Marsden *et al*., 2014). Our results show strong differentiation within and between populations, but find these allelic differences are only seen when susceptible and resistance replicates are compared (with no resistant *versus* resistant or susceptible *versus* susceptible signal). The loci responsible for insecticide resistance in this region may only be found on one arrangement of the 2Rc inversion, but the resolution of the pool-GWAS was still much finer than the length of inversion (in *A. gambiae* >5Mb - White *et al*., 2009), with replicates and multiple independent population comparisons narrowing the candidate to just ~225kb.

5.5.2 Candidate genes

As part of the ongoing *Anopheles* 16 Genomes Project (Neafsey *et al*., 2015), families of known detoxification genes were manually annotated on the *A. arabiensis* reference genome. Available from VectorBase for the AraD1 reference genome (Giraldo-Calderón *et al*., 2014),

these annotations allowed us to investigate potential gene- and SNP-level drivers of candidate regions identified with by pool-GWAS. The fully concordant candidate region on the 2R chromosome arm, overlapped a cluster annotated cytochrome P450 genes from the CYP6 family, four genes actually within the candidate region and five more genes from this family just outside of the region (included in the extended region candidate SNP analysis) (Appendix 5.7.3). This family of genes are strong candidates for involvement in the breakdown of xenobiotics, with *A. gambiae Cyp6p3* and *Cyp6p4* (which lies within the candidate region) demonstrating metabolism of type 1 and 2 pyrethroids using mosquito genes expressed in *Escheria coli* membranes (Müller *et al*., 2008; Mark Paine pers. comm.).

Confidence in this CYP gene cluster, falling within the candidate region association with a resistant phenotype, is strengthened further by it falling within a region linked with pyrethroid resistance from a recent QTL study on *A. arabiensis* in Chad. Witzig and colleagues (2013), using genetic crosses between resistant and susceptible strains, identified a single large QTL ~14Mb in size which explained 24.4% of the insecticide resistance variance found. Due to the nature of QTL study it was not possible to increase genomic resolution further, this being limited by recombination and the study's use of F2 backcrosses, hence the QTL was large (increasing the number of generations would be necessary to break up the QTL). However, of the six genes (all CYP450s) the authors followed up with expression studies, the only one with a significant signal (>20x overexpression in resistant mosquitoes) was *Cyp6p4*, a gene in our much smaller ~225kb candidate region (Witzig *et al*., 2013). The same gene was also found upregulated in resistant *A. arabiensis* from Uganda (Currie-Jordan, 2015), suggesting the same gene may have similar roles across different resistant *A. arabiensis* populations.

Though *Cyp6p4* appeared a strong hit in the candidate region, there are several genes in the cluster, covered by the extended candidate region (extended to allow for window edge effects and detection of gene regulators), we identify that have not yet been investigated. Alongside the CYP450s (that we were able to investigate due to manual annotation) there are also other genes of interest inside or close to the extended candidate region including a carboxylesterase (COE) alpha esterase (AARA015786) the family of which have known involvement in insecticide resistance (Ranson *et al*., 2002) and a gene (AARA004675) with an orthologue of a sodium-calcium exchange protein in *Drosophila melanogaster*. Though sodium-calcium exchange proteins are not usually associated with insecticide resistance, they have been found upregulated in insecticide resistant *A. gambiae* (Vontas *et al*., 2005)

and selective sweep signature has been found centred upon a gene from this family in the *Anopheles gambiae* 1000 Genomes Project data (Alistair Miles pers. comm.). With the pool-GWAS reducing the candidate region to one spanning so few genes, an investigation of all of them using either qPCR for potential over expression, targeted individual sequencing to look for regulatory/coding mutations or *in-vitro* methods to test for metabolism is now both warranted and tractable.

5.5.3 Candidate SNPs

Manual annotation, ensuring the correct start and end points of coding regions, enabled insecticide resistance association analysis at an even higher resolution, at the SNP level. Though previous studies, particularly those investigating CYP450s had found gene over-expression as the mode of insecticide resistance, it should be noted that generally the genes were not sequenced (Witzig *et al*., 2013) or that few individuals were sequenced (Currie Jordan, 2015). Evidence from *A. gambiae* insecticide resistance also indicates that apparent over expression can be due to copy number variation of resistance point mutation carrying and susceptible copies of duplicated *Ace-1* genes (Djogbénou *et al*., 2008b; Weetman *et al*., 2015). Analysis of SNPs outside of coding regions may also allow location of any *cis* regulatory drivers of potential expression changes. However, no resistant *versus* susceptible (or Moshi alive *versus* dead) highly divergent SNPs were found that were concordant across all comparisons, with just Moshi dead *versus* Tarime producing a candidate non-synonymous stop-lost SNP in *Cyp6p15p*, just ~7kb outside the candidate region, a gene which, to our knowledge, has not previously been investigated, and now should be followed up with functional analysis.

Several hypotheses may explain the strongly resistance associated candidate region with a lack of an apparent hard selective sweep. One that we were able to test within the scope of this study was gene duplication; as demonstrated by the *Ace-1* gene mediated resistance in *A. gambiae* (Weetman *et al*., 2015). Gene duplication leading to CNVs can alter allele frequencies without hard selective sweep signature (Pritchard, Pickrell and Coop, 2010; Messer and Petrov, 2013). In WGS/pool-seq the depth of sequencing coverage can indicate recently duplicated regions (with high similarity to ancestral sequence) as having approximately twice or more times the coverage of the surrounding region (Yoon *et al*., 2009). However the 2R candidate region shows no concordant increase in sequencing depth which suggested this is not the case here. Another hypothesis for apparent lack of hard

selective sweep is one of multiple soft sweeps occurring on a locus or loci within the candidate region, obscuring a concordant hit across all comparisons. Soft sweeps may be driven by insecticidal resistance evolving from standing variation, upon population exposure to insecticides multiple loci with fitness improvements increase in frequency (Messer and Petrov, 2013). Overlaying of selective sweeps, as found with the *Vgsc-1575Y* mutation further increasing pyrethroid insecticide resistance of *Vgsc-1014F kdr* haplotype may also have the same hard sweep obscuring effects (Jones *et al*., 2012b). Detection of several high frequency haplotypes associated with resistance could suggest soft or overlaid sweeps (Jones *et al*., 2012b; Messer and Petrov, 2013), though haplotype estimation is difficult with pool-seq data, as (in our data) 80 haplotypes potentially exist within each pool (of 40 individuals) and as the sequencing read length is short (100bp), LD may be hard to detect (Schlötterer *et al*., 2014). Fortunately several algorithms have been developed to try to estimate haplotypes *de novo* (with no prior haplotype knowledge) from pool-seq data (Iliadis, Anastassiou and Wang, 2012; Cao and Sun, 2014), including one proven to be accurate using pools containing many individuals (Long *et al*., 2011).

Another possible reason for the lack of concordant SNPs in the insecticide resistance candidate region on the 2R chromosome arm is the direction of the analysis we conducted. The mainland inter-population comparison contrasts Moshi dead with Tarime allele frequencies, rather than using Moshi alive *versus* Tarime. The rationale behind this choice was that with both alive and dead coming from the same resistant population (with the dead individuals not surviving insecticide bioassay), detection of resistant (Moshi dead) *versus* susceptible population (Tarime) allele frequencies would allow an independence of expectation (Tarime < Moshi dead < Moshi alive) in the absence of complete independence in data (the necessity of using a Moshi group twice). Though this logic held true at the candidate region level, and with 1000 variants contributing to each window divergent allele frequency signals were clearly seen and concordant regions defined, at the individual SNP level this approach relies on enough of the resistance associated variant to be present in the Moshi dead pools to pass filters.

5.5.4 *A. merus* as an outgroup for *A. arabiensis*

Unfortunately using *A. merus* as an outgroup for *A. arabiensis*, it being fixed for inversions on 2R rather than polymorphic (Coluzzi *et al*., 2002), was unsuccessful due to low density of markers. However, with a recent study estimating 1.85 ± 0.47 million years separating the

two species and for the first time quantifying the benefits of aligning WGS *Anopheles* species to their (recently released) respective reference genomes (Fontaine *et al*., 2015), our result is perhaps unsurprising. With such ancient divergence and indeed low initial sequencing coverage of the *A. merus* pool, a relaxing of the filtering rules by reducing the subsampling threshold (as per the Zanzibar pools) may increase the resolution of available variants. *In silico* methods are also being developed for karyotyping WGS individuals (*Anopheles gambiae* 1000 Genomes Project – unpublished) and it may be possible to extend these techniques to pool-seq data, bypassing the need to compare to other species with fixed inversion arrangements to elucidate karytopes.

5.5.5 *Cyp4g16*

The 2R insecticide resistance candidate region highlighted in this pool-GWAS contains a cluster of CYP450 genes from the 6 family, some of which are known to metabolise pyrethroids in Anophelines (Müller *et al*., 2008) and have been associated with resistant phenotypes in *A. arabiensis* (Witzig *et al*., 2013). Another CYP450 gene that has been repeatedly implicated in *A. arabiensis,* through upregulation in resistant individuals, is *Cyp4g16* (Jones *et al*., 2013; Matowo *et al*., 2014b). However, the contig upon which this gene is found in AraD1 reference genome was unable to be mapped to the chromosome arm assembly generated using orthology (AraChr - Sharakhov, Jiang and Hall, unpublished), due to not carrying enough genes, and any association between this gene and insecticide resistance could not be detected in our sliding window analysis. A targeted approach was, therefore, taken where the all SNPs on the *Cyp4g16* carrying contig were examined. However, a lack of concordantly resistant associated SNPs not just between but also within comparisons suggests that no single variant sequenced on the contig is linked with a resistant phenotype. One possible explanation, as discussed with reference to the 2R candidate region, is that multiple soft sweeps may be obscuring signatures of insecticidal selection on the region (Messer and Petrov, 2013) or possibly the upregulation of the gene found previously (Jones *et al*., 2013; Matowo *et al*., 2014b) is *trans* regulated by a genomic region not found on the small contig containing *Cyp4g16*. Alternatively, resistance associated variants may be present in the region but were not sequenced. A ~5000bp gap in sequencing coverage was found on the contig, possibly due to difficulty in sequencing or aligning that region. Re-sequencing with longer read technology such as PacBio (Ferrarini *et al*., 2013), may help cover these gaps in the future.

5.5.6 Future work

The alignment of sequenced reads to a version of the *A. arabiensis* reference genome assembled in chromosome arms, five arms rather than 1214 contigs, enabled a sliding window approach to be employed. However, by nature of the syntenic method of production: a contig must have five orthologous genes in synteny with *A. gambiae* AgamP3 reference genome to be placed on chromosome arm - smaller contigs have not been considered in our main analyses. Though the mean contig size in the AraD1 reference genome is 203103.7 nucleotides, the distribution is highly positively skewed with a median contig length of just 9896 nucleotides, too small to carry the genes needed for syntenic matching which lead to the vast majority of the contigs being left out of the analyses. This necessitated the targeted separate targeted analysis of the KB704462 contig, missing from the chromosome arm reference due to size (22284bp) to follow up *Cyp4g16*, a candidate gene from previous research (Jones *et al*., 2013; Matowo *et al*., 2014b). It should be noted that despite these smaller contigs not being considered, 87.7% of the AraD1 reference has been successfully aligned to chromosome arms (Appendix 5.7.5). A targeted approach could also be taken with other genes shown to be associated with pyrethroid resistance in other insects, such as candidates from microarray studies (*e.g.* Bariami *et al*., 2012). Future work should involve improvement of the reference genome to include the missing 12.3% of small but potentially important genomic regions, either by relaxing the number of syntenic genes required for placement or preferably by re-sequencing *A. arabiensis* using longer read technology such as Pac-Bio (Ferrarini *et al*., 2013) or Illumina TruSeq (McCoy *et al*., 2014). The reference genome would also be further improved by annotation of features such as genes and inversions, allowing accurate association with phenotypes. In many cases this will require manual annotation, however, which given the number of genes (there are 12,843 coding genes in *A. gambiae* AgamP4.2 gene build (Giraldo-Calderón *et al*., 2014)), will be an arduous task requiring scientific community participation.

The analysis here demonstrates that resistance candidates can be identified using a pool-seq GWAS approach in *A. arabiensis*. However, only a single candidate region was explored in detail here; on the mainland (Moshi/Tarime) there was another strong candidate on 3L and multiple candidates across the island comparisons (Unguja/Pemba). The next steps would be to investigate these in detail, potentially using the Moshi alive *versus* Tarime analysis instead of Moshi dead, to identify candidate genes and SNPs then follow up with molecular analysis to validate. A Moshi alive *versus* Tarime comparison may increase the chances of finding

SNPs associated with resistance that are concordant across comparison, however, these variants may not be detectable due to soft selective sweeps (Messer and Petrov, 2013) so a germane analysis would be to utilise algorithms designed to detect these sweeps in pooled data (Long et al., 2011; Iliadis, Anastassiou and Wang, 2012; Cao and Sun, 2014). Sequencing coverage analysis also identified potential CNVs and appeared concordant between geographically separated populations. Though no link with these peaks of high coverage and insecticide driven directional selection was found they could be investigated to determine if these are driven by selfish elements (Li *et al*., 2005), genomic structure (Coluzzi *et al*., 2002) or perhaps selection on another phenotype not tested here.

5.5.7 Conclusion

Through the design of a study with both internal replication and independent comparisons we have demonstrated how pool-GWAS can be successfully utilised in *A. arabiensis* to confidently locate regions of the genome associated with an insecticide resistance phenotype. For the 2R candidate region explored in detail here, further corroboration of involvement in resistance comes from the region falling within a previously established resistance QTL from a completely separate population (Witzig *et al,*. 2013) and because it contains genes, the over expression of which has been linked with a pyrethroid resistance phenotype in the species (Jones *et al*., 2013; Matowo *et al*., 2014b). We were unable to isolate SNPs in the candidate region with concordant resistance association across all comparisons, potentially due to the conservative nature of our analysis or perhaps due to the effect of soft sweeps (Messer and Petrov, 2013), however a candidate SNP for mainland resistance was established, affecting the function of *Cyp6p15p*, a gene which now requires further investigation. With a wealth of data produced and other resistance candidates to explore, the pool-GWAS technique has proven to be a promising technique for establishing phenotype-genotype associations in malaria vector mosquitoes.

5.6 Acknowledgments

Many thanks to Igor Sharakhov, Xiaofang Jiang and Brantley Hall, who aligned the reference genome contigs to chromosome arms, making our windowed analytical approach possible and to Bilali Kabula and Chris Jones for samples.

5.7 Appendix

Appendix 5.7.1 Pool information

Appendix 5.7.1. Table of pool information. Collection site, pyrethroid insecticide resistance phenotype and population type are shown for each pool with mean sequencing coverage calculated from .bam files.

Pool	Site	Phenotype	Population type	Mean depth
Anopheles_arabiensis_Tz_1	Lower Moshi	alive	res	129.7
Anopheles_arabiensis_Tz_2	Lower Moshi	alive	res	123.7
Anopheles_arabiensis_Tz_3	Lower Moshi	alive	res	141.1
Anopheles_arabiensis_Tz_4	Lower Moshi	alive	res	127.9
Anopheles_arabiensis_Tz_5	Lower Moshi	alive	res	118.4
Anopheles_arabiensis_Tz_6	Lower Moshi	alive	res	130.8
Anopheles_arabiensis_Tz_7	Lower Moshi	alive	res	124.8
Anopheles_arabiensis_Tz_8	Lower Moshi	alive	res	162.3
Anopheles_arabiensis_Tz_9	Lower Moshi	alive	res	155.9
Anopheles_arabiensis_Tz_10	Lower Moshi	alive	res	165.3
Anopheles_arabiensis_Tz_11	Lower Moshi	dead	res	122.0
Anopheles_arabiensis_Tz_12	Lower Moshi	dead	res	137.1
Anopheles_arabiensis_Tz_13	Lower Moshi	dead	res	135.9
Anopheles_arabiensis_Tz_14	Lower Moshi	dead	res	143.2
Anopheles_arabiensis_Tz_15	Lower Moshi	dead	res	145.7
Anopheles_arabiensis_Tz_16	Lower Moshi	dead	res	151.7
Anopheles_arabiensis_Tz_17	Tarime	sus	sus	129.7
Anopheles_arabiensis_Tz_18	Tarime	sus	sus	158.8
Anopheles_arabiensis_Tz_19	Tarime	sus	sus	138.0
Anopheles_arabiensis_Tz_20	Tarime	sus	sus	156.1
Anopheles_arabiensis_Tz_21	Tarime	sus	sus	162.8
Anopheles_arabiensis_Tz_22	Tarime	sus	sus	140.6
Anopheles_arabiensis_Tz_23	Unguja	sus	sus	72.2
Anopheles_arabiensis_Tz_24	Unguja	sus	sus	60.3
Anopheles_arabiensis_Tz_25	Pemba	res	res	147.2
Anopheles_arabiensis_Tz_26	Pemba	res	res	139.1
Anopheles_merus_Tz_1	Pemba	unknown	unknown	58.0

Appendix 5.7.2 Commands

Command 1.

picard-tools MarkDuplicates INPUT=<inFile.bam> OUTPUT=<outFile.bam> METRICS_FILE=<outFile> REMOVE_DUPLICATES=true ASSUME_SORTED=true

Command 2.

picard-tools MergeSamFiles INPUT=<inFile.bam> INPUT=<inFile.bam> INPUT=… OUTPUT=<mergedOut.bam> ASSUME_SORTED=false USE_THREADING=true MERGE_SEQUENCE_DICTIONARIES=true

Command 3.

picard-tools SamToFastq INPUT=<inFile.bam> FASTQ=<outFile1.fastq> SECOND_END_FASTQ=<outFile2.fastq> INTERLEAVE=false

Command 4.

bwa sampe -r '@RG\tID:<ID>\tPL:illumina\tLB:<library>\tSM:<sample>' <refFasta> <sai1> <sai2> <fastq1> <fastq2> > <outFile.sam>

Command 5.

samtools view –q 20 –bS <inFile.sam> | samtools sort - <outFile.bam

Command 6.

genomeCoverageBed -ibam <inFile.bam> -g <genome.txt> > <outFile.coverage>

Command 7.

samtools mpileup –B <inFile1.bam> <inFile2.bam> <inFile*n*.bam> > <outFile.mpileup>

Command 8.

java –Xmx10g –jar mpileup2sync.jar --input <inFile.mpileup> --output <outFile.sync> --fastq-type sanger –min-qual 20 –threads 8

Command 9.

subsample-synchronised.pl –input <inFile.sync> --output <outFile.sync> --target-coverage 80 –max-coverage 2% --method withoutreplace

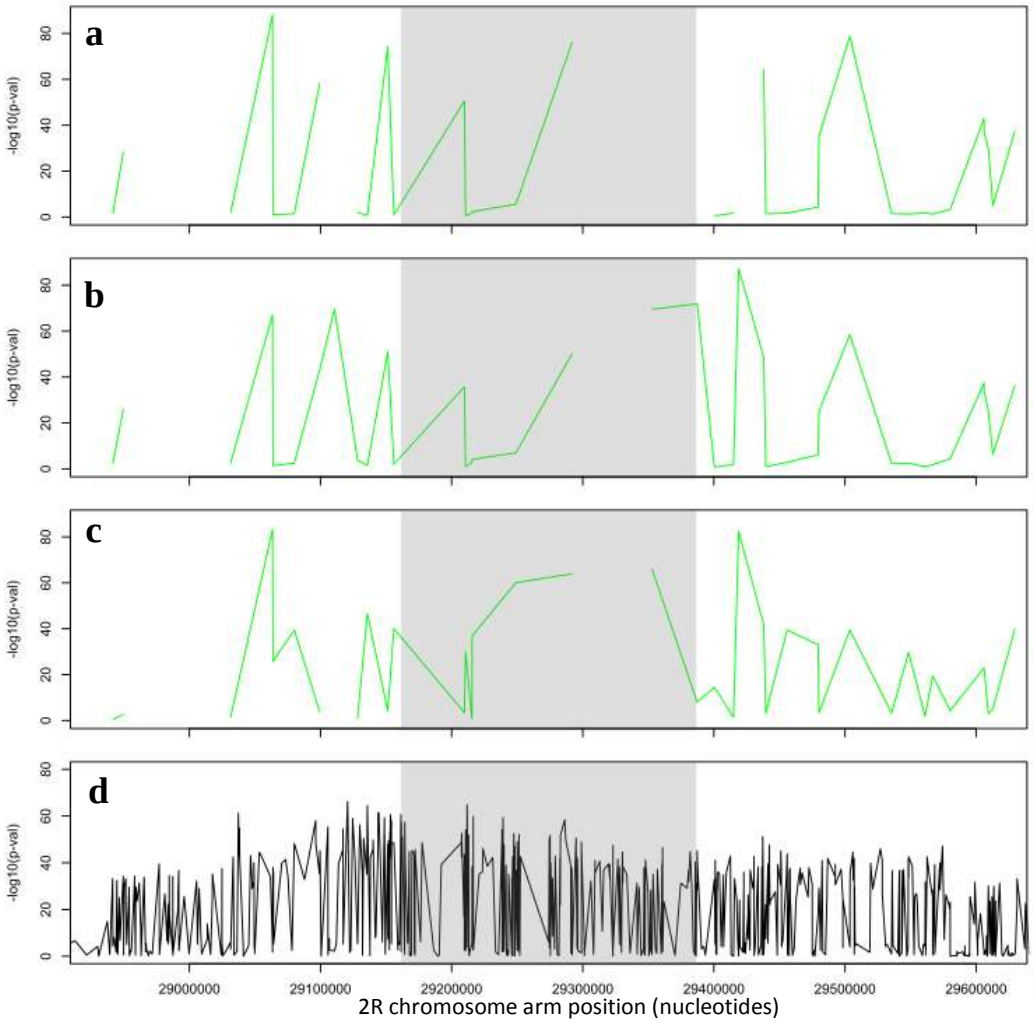
Appendix 5.7.3 Genes in candidate region.

Appendix 5.7.3. Genes found within 2R candidate region. Genes which lie within the cross comparison insecticide resistance candidate region and extended candidate region were extracted from the AraD1.3 *A. arabiensis* gene-set using VectorBase (Giraldo-Calderón *et al*., 2014). Positions transposed from AraChr to AraD1. Extended region KB704451: 10581961- 11032311, candidate region KB704451:10807136-11032311. Gene names are given for the manually annotated genes.

Gene stable ID	Gene name	Region
AARA004665		Extended
AARA004666		Extended
AARA004667		Extended
AARA004668		Extended
AARA004669		Extended
AARA004670		extended
AARA004671		extended
AARA004672		extended
AARA004673		extended
AARA004674		extended
AARA004675		extended
AARA004676	CYP6AA1	extended
AARA014411		extended
AARA004677	CYP6AA2	extended
AARA015786		extended
AARA015785	CYP6P15P	extended
AARA015787	CYP6P3	extended
AARA015788	CYP6P5	extended
AARA015789	CYP6P4	candidate
AARA015790	CYP6P1	candidate
AARA015791	CYP6P2	candidate
AARA015792	CYP6AD1	candidate
AARA004679		candidate
AARA004680		candidate
AARA004681		candidate
AARA004682		candidate
AARA004683		candidate

AARA004684	candidate
AARA004685	candidate
AARA004686	candidate
AARA004687	candidate
AARA004688	candidate
AARA004689	candidate
AARA004690	candidate
AARA004691	candidate
AARA004692	candidate
AARA004693	candidate
AARA004694	candidate
AARA004695	candidate
AARA004696	candidate
AARA014412	extended
AARA004697	extended
AARA004698	extended
AARA004699	extended
AARA004700	extended
AARA004701	extended
AARA004702	extended
AARA004703	extended
AARA014413	extended

Appendix 5.7.4 *A. merus* inversion comparison



Appendix 5.7.4. Pairwise 2R chromosome arm p-values for *A. arabiensis* comparisons with *A. merus* – extended candidate region. (**a**) Moshi alive *versus A. merus*. (**b**) Moshi dead *versus A. merus*. (**c**) Tarime *versus A. merus*. (**d**) For comparison, Moshi dead *versus* Tarime. The shaded region highlights the insecticide resistance candidate region concordant across all comparisons. Interspecies comparisons are shown in green, intra-species in black.

Appendix 5.7.5 Chromosome sizes

Appendix 5.7.5. Number and sizes (chromosome arms only) of reference genome contigs. The AraChr reference being constructed by assigning the AraD1 *A. arabiensis* reference genome contigs to chromosome arms using orthologue synteny with the *A. gambiae* reference genome – AgamP3 (Holt *et al*., 2002).

Reference	Contigs	2L bp	2R bp	3L bp	3R bp	X bp	Total bp
AraD1	1214	-	-	-	-	-	246567867
AraChr	5	47444456	59049218	38873271	49805857	21162472	216335274
AgamP3	5	49364325	61545105	41963435	53200684	24393108	230466657

Chapter 6

Final discussions and conclusions

6.1 Discussion

Improvements in genome sequencing technology and rapidly decreasing costs (Baker, 2010) have seen whole genome sequence (WGS) data applied to manifold evolutionary questions, across a plethora of systems, generating new and adapting old population genetics methodologies to these vast, high resolution data sets (Seehausen *et al*., 2014). *Anopheles* malaria vectors provide tractable systems for study using genomic techniques due to easy access to powerful resources (*e.g.* VectorBase – Giraldo-Calderón *et al*., 2014; 16 Genomes Project – Neafsey *et al*., 2015 and the *Anopheles gambiae* 1000 Genomes Project (Ag1000G) - https://www.malariagen.net/) and because over half a century of research has resulted in a wealth of phenotypic and other 'meta' data being collected for many populations. With insecticide campaigns generating high levels of directional selection (Lynd *et al*., 2010), malaria vectors offer a great opportunity to investigate how species adapt in fast changing ecology. Although populations have been studied at the molecular level for a number of years, in this thesis genomic techniques were explored to gain new insights into these mosquitoes both from an evolutionary perspective, elucidating how these animals diverge in the face of differing levels of gene flow or how they adapt quickly to changing environments, and from a medical perspective, for example how does insecticide resistance evolve and are we able to detect the loci behind it.

6.1.1 Chapter 2 – Ghana *kdr* introgression

In this chapter we carry out the first WGS study using wild populations of *A. gambiae* (S-form) and *A. coluzzii* (M-form). Samples with and without knock down resistance (*kdr*) mutations were compared to investigate the introgression of an insecticide resistance locus on a genomic scale. With small sample sizes, a concern when designing the experiment was that there may be a lack of power to establish confidence in any signals found. However, results revealed that our pairwise comparisons of just five *versus* five individuals were quite capable of detecting and quantifying the striking signals of divergence found between these samples. Confidence in the signals found was gained by the concordance across few samples

but over large numbers (thousands) of SNPs, a level of marker density only viable with WGS. These results suggest that future WGS projects on *Anopheles* can be cost effective with small numbers of individual sequences. The adaptive introgression of the resistance conferring locus, *Vgsc-1014F,* from *A. gambiae* in to *A. coluzzii,* had transferred a ~3Mb region of the genome with it. With no discernible impact on reproductive isolation, data also suggest that these species are resilient to even large gene flow events across their demonstrably porous species barrier and highlighted how introgression can allow fast evolutionary adaptive responses to anthropogenic environmental changes, an important factor when designing vector control approaches.

Recent research revived the debate on the use of relative measures of divergence *versus* the use of absolute measures. Relative measures, like the commonly used F_{ST}, take into account the diversity within populations as well as between, if divergence within populations is low then F_{ST} values between them can be inflated (Charlesworth, Nordberg and Charlesworth, 1997; Charlesworth, 1998; Cruickshank and Hahn, 2014). These effects may be problematic when considering genome scans for divergence, for example the approach we took with *A. gambiae* and *A. coluzzii* in Chapter 2, as reduced recombination around centromeres can leave regions depauperate of diversity (Charlesworth, 1998) inflating divergence. To combat these problems the use of an absolute measure of divergence, D_{xy} (Takahata and Nei, 1985), which is not affected by within population diversity has been suggested (Cruickshank and Hahn, 2014). We used both absolute and relative measures but also found problems of high variance with D_{xy} indicating that it may not be a useful replacement for F_{ST} (Wakeley, 1996). The results suggest that with no 'perfect' statistic, when conducting genome-wide divergence scans it may be advisable to take a holistic approach, using orthogonal measures.

6.1.2 Chapter 3 – Genomic replacement in Guinea Bissau

The former molecular forms of the major malaria vector *A. gambiae* (M and S) are now thought to be recently diverged sibling species and were elevated to specific status, S-form keeping the original nomenclature (*A. gambiae*) and M-form being renamed *A. coluzzii* (Coetzee *et al.*, 2013). Much of the justification for this status change came from molecular data suggesting gene flow across most of the mosquitoes' range was low (della Torre, Tu and Petrarca, 2005, Simard *et al.,* 2009; Tripet *et al.,* 2001) and genomic work suggesting genome wide divergence between the species pair (Lawniczak 2010; Reidenbach 2012). However, in the far-west of the range, high gene-flow has been recorded (Caputo *et al.,* 2008;

Oliveira *et al.,* 2008) and this chapter discusses genomic data used to augment microsatellite data to investigate this 'aberrant' gene-flow scenario. We find that rather than being distinct reproductively isolated units, in the coastal region of Guinea Bissau these units are breaking down with asymmetric introgression from *A. coluzzii* into *A. gambiae*. The results question the validity of elevation to species status when a large range of gene flow is found. Though it may be helpful, scientifically, to partition these mosquitoes into separate species in other regions, in high gene flow regions like Guinea Bissau it may not.

Though dropping in price (Baker, 2010), the high resolution WGS data is much more expensive to produce than the lower resolution microsatellite techniques. Financial constraint meant that WGS data was only available for three of the eight sample locations covered by the microsatellite data. However, techniques developed in the Ghanaian introgression work of Chapter 2 were able to be employed by using resources and data developed in other *A. gambiae* projects. Comparison of samples from Guinea Bissau to a 'control' population, with much lower gene, flow was possible due to genome sequences from Ghana, generated by our earlier work (Clarkson *et al*., 2014) and though we did not have WGS for *A. coluzzii* from Guinea Bissau, we were able to interrogate genome-wide ancestry informative makers from the species using markers generated from research by Neafsey *et al.* (2010). The insight gained into the genomic landscape of high gene flow between these closely related malaria vectors would not have been possible without these genomic resources being in the public domain and with both the 16 Genomes Project (Neafsey *et al*., 2015) and Ag1000G (https://www.malariagen.net/) making a wealth of new genomic resources available, question can be asked on the genomic level across much of these mosquitoes range, encompassing many medically relevant environmental and genetic clines.

6.1.3 Chapter 4 – Voltage gated sodium channel gene networks

As sequencing technologies improve, ever larger genomic data set amass, with analysis and visualisation techniques essential to exploiting this data struggling to keep pace (Krzywinski *et al*., 2009). In Chapter 4 we investigated the DNA sequence of a gene involved in insecticide resistance, the voltage gated sodium channel (VGSC), using a massive WGS data set composed of hundreds of individuals. We demonstrate that a haplotype network approach, a technique often utilised to visualise relationships between much smaller numbers and diversity of haplotypes (*e.g.* Pinto *et al.*, 2007; Etang *et al*., 2009), is capable of 'scaling up' and still produce intuitive and visually informative figures. By constructing a network, the

evolutionary relationships, in terms of diversity and divergence, between the 1530 haplotypes were clear to see. The software used, Cytoscape 3.1 (Smoot *et al.* 2010), allowed the visual layering of additional data upon the network; it was this feature which clearly elucidated features that otherwise may have gone unnoticed in the data, the high levels of long distance haplotype sharing between populations or the high concentrations of non-synonymous mutations in certain regions of the network, for example. With whole genome haplotyping underway for all individual sequences in the Ag1000G and with many genes (or other genomic regions) still of interest in *A. gambiae*, our results show that a powerful, hypothesis generating technique not only exists to visualise the data, but by using Cytoscape, is simple to use and freely available.

6.1.4 Chapter 5 – Insecticide resistance in *A. arabiensis*, a GWAS approach

A recent study researching the viability of association studies in the malaria vector mosquito *A. arabiensis* suggested that future work would require large sample sizes to overcome the low levels of linkage disequilibrium (Marsden *et al*., 2014). Large sample sizes often meaning the sequencing of hundreds of individual genomes to gain confidence in genome wide association study (GWAS) candidates (Manolio *et al*., 2009; Park *et al*., 2010), but the cost of this can be prohibitive for most studies. This is unfortunate as association studies are a proven technique for linking phenotypes with genotypes, in malaria vectors an attribute that could be put to use to discover the loci and mechanisms driving insecticide resistance, a major threat to vector control campaigns (World Health Organization, 2012). To overcome this sample size constraint of traditional WGS GWAS, we designed a study using a pool-seq approach, where multiple individuals are sequenced together. This allowed us to perform a GWAS using over 1000 individuals in 27 pools for a fraction of the cost of individual WGS that many individuals. To our knowledge pool-GWAS has not be utilised on *Anopheles* before and we demonstrated its viability as a technique for detecting regions of the genome associated with an insecticide resistance phenotype. Unfortunately the technique we used can only be replicated on *A. gambiae* and *A. arabiensis* due to them both having reference genomes assembled into long chromosome arm contigs, however, with several other vector anophelines recently gaining reference genomes (Neafsey *et al*., 2015) and techniques being developed to generate chromosome alignments from these (Sharakhov, Jiang and Hall, unpublished), pool-GWAS should prove to be relatively cheap and powerful tool for investigation of adaptation across other malaria vectors too.

6.2 Conclusion

The results contained within this thesis demonstrate the power of and applications for genomics in furthering our understanding, both specifically of these medically important animals, and more generally of the evolutionary processes driving adaption and speciation. The sibling species of *A. gambiae* and *A. coluzzii* have established themselves as a natural system for the study of speciation with Turner, Hahn and Nuhzdin's seminal paper (2005). However, the recent increase in availability of high quality reference genomes and other community curated genomic resources may see *Anopheles* begin to rival *Drosophila* and humans as model systems for the study of a much broader range of evolutionary questions, while our increased understanding of these vectors takes us closer to malaria's eradication.

7.1 References

1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." *Nature* 467.7319 (2010): 1061-1073.

Aboagye-Antwi, F. *et al.* "Experimental swap of *Anopheles gambiae*'s assortative mating preferences demonstrates key role of X-chromosome divergence island in incipient sympatric speciation." *PloS Genetics* (2015): e1005141.

Andrew, R. L., and L. H. Rieseberg. "Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes." *Evolution* 67 (2013): 2468-2482.

Asidi, A. *et al.* "Loss of household protection from use of insecticide-treated nets against pyrethroid-resistant mosquitoes, Benin." *Emerging Infectious Diseases* 18.7 (2012): 1101.

Athrey, G. *et al.* "The effective population size of malaria mosquitoes: large impact of vector control." *PLoS Genetics* 8.12 (2012).

Awolola T. S. *et al.* "Evidence of multiple pyrethroid resistance mechanisms in the malaria vector *Anopheles gambiae* sensu stricto from Nigeria." *Transactions of the Royal Society of Tropical Medicine* 103 (2009): 1139-1145.

Bagi, J. *et al.* "When a discriminating dose assay is not enough: measuring the intensity of insecticide resistance in malaria vectors." *Malaria Journal* 14.1 (2015): 210.

Baird, N. A. *et al.* "Rapid SNP discovery and genetic mapping using sequenced RAD markers." *PloS one* 3.10 (2008).

Baker, M. "Next-generation sequencing: adjusting to data overload." *Nature Methods* 7.7 (2010): 495-499.

Barnes, M. J. *et al.* "SINE insertion polymorphism on the X chromosome differentiates *Anopheles gambiae* molecular forms." *Insect Molecular Biology* 14.4 (2005): 353-363.

Barrett, J.C. *et al.* "Haploview: analysis and visualization of LD and haplotype maps." *Bioinformatics* 21.2 (2005): 263-265.

Barton, N. H. and B. Charlesworth. "Genetic revolutions, founder effects, and speciation." *Annual Review of Ecology and Systematics* (1984): 133-164.

Bass, C. *et al.* "Detection of knockdown resistance (*kdr*) mutations in *Anopheles gambiae*: a comparison of two new high-throughput assays with existing methods." *Malaria Journal* 6, 111 (2007).

Bastide, H. *et al*. "A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*." *PLoS Genetics* (2013): e1003534.

Bayoh, M. N. *et al*. "*Anopheles gambiae*: historical population decline associated with regional distribution of insecticide-treated bed nets in western Nyanza Province, Kenya." *Malaria Journal* 9.1 (2010): 62.

Bhatt, S. *et al*. "The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015." *Nature* 526.7572 (2015): 207-211.

van den Berg, H. *et al*. "Global trends in the use of insecticides to control vector-borne diseases." *Environmental Health Perspectives* 120.4 (2012): 577-582.

Besansky, N. J. *et al*. "Semipermeable species boundaries between *Anopheles gambiae* and *Anopheles arabiensis*: evidence from multilocus DNA sequence variation." *Proceedings of the National Academy of Sciences* 100.19 (2003): 10818-10823.

Bolnick, D. I. and B. M. Fitzpatrick. "Sympatric speciation: models and empirical evidence." *Annual Review of Ecology, Evolution, and Systematics* (2007): 459-487.

Brito, L. P. *et al*. "Assessing the effects of *Aedes aegypti kdr* mutations on pyrethroid resistance and its fitness cost." *PLoS One* 8.4 (2013): e60878.

Brooke, B. D., R. H. Hunt, and M. Coetzee. "Resistance to dieldrin+ fipronil assorts with chromosome inversion 2La in the malaria vector *Anopheles gambiae*." *Medical and Veterinary Entomology* 14.2 (2000): 190-194.

Brooke, B. D. *et al*. "Bioassay and biochemical analyses of insecticide resistance in southern African *Anopheles funestus* (Diptera: Culicidae)." *Bulletin of Entomological Research* 91.04 (2001): 265-272.

Bulmer, M. G. "*Principles of Statistics"* (Dover, New York, 3rd Ed, 1979).

Burton, M. J. *et al*. "Differential resistance of insect sodium channels with kdr mutations to deltamethrin, permethrin and DDT." *Insect Biochemistry and Molecular Biology* 41.9 (2011): 723-732.

Butlin, R. K. "Recombination and speciation." *Molecular Ecology* 14.9 (2005): 2621-2635.

Butlin, R. and C. Roper. "Evolutionary genetics: microarrays and species origins." *Nature* 437.7056 (2005b): 199-201.

Butlin, R. K., J. Galindo and J. W. Grahame. "Sympatric, parapatric or allopatric: the most important way to classify speciation?" *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1506 (2008): 2997-3007.

Cao, C-C. and X. Sun. "Accurate estimation of haplotype frequency from pooled sequencing data and cost-effective identification of rare haplotype carriers by overlapping pool sequencing." *Bioinformatics* (2014).

Caputo, B. *et al*. "*Anopheles gambiae* complex along The Gambia river, with particular reference to the molecular forms of *An. gambiae* ss." *Malaria Journal* 7.1 (2008): 182.

Carneiro, M., N. Ferrand and M. W. Nachman. Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics* 181(2008): 593-606.

Cassone, B. J. *et al*. "Gene expression divergence between malaria vector sibling species *Anopheles gambiae* and *An. coluzzii* from rural and urban Yaounde Cameroon." *Molecular Ecology* 23.9 (2014): 2242-2259.

Chandre, F. *et al*. "Status of pyrethroid resistance in *Anopheles gambiae sensu lato*." *Bulletin of the World Health Organization* 77.3 (1999): 230-234.

Chang C. C. *et al*. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4.

Charlesworth, B. "Background selection and patterns of genetic diversity in Drosophila melanogaster." *Genetical Research* 68.2 (1996): 131-150.

Charlesworth, B., M. Nordborg and D. Charlesworth. "The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations." *Genetics Research* 70 (1997): 155-174.

Charlesworth, B. "Measures of divergence between populations and the effect of forces that reduce variability." *Molecular biology and evolution* 15.5 (1998): 538-543.

Chevin, L. M., S. Billiard and F. Hospital. "Hitchhiking both ways: effect of two interfering selective sweeps on linked neutral variation." *Genetics* 180 (2008): 301-316.

Chinery, W. A. "Effects of ecological changes on the malaria vectors *Anopheles funestus* and the *Anopheles gambiae* complex of mosquitoes in Accra, Ghana." *The Journal of Tropical Medicine and Hygiene* 87.2 (1984): 75-81.

Chiu, T-L *et al*. "Comparative molecular modeling of *Anopheles gambiae* CYP6Z1, a mosquito P450 capable of metabolizing DDT." *Proceedings of the National Academy of Sciences* 105.26 (2008): 8855-8860.

Cingolani, P. *et al*. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3." *Fly* 6.2 (2012): 80-92.

Clarkson, C. S. *et al*. "Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation." *Nature Communications* 5 (2014).

Clement M, D. Posada D and K. Crandall. "TCS: a computer program to estimate gene genealogies." *Molecular Ecology* 9(10) (2000): 1657-1660.

Coetzee, M. *et al*. "*Anopheles coluzzii* and *Anopheles amharicus*, new members of the Anopheles gambiae complex." *Zootaxa* 3619.3 (2013): 246-274.

Collins, F. H. *et al*. "A Ribosomal RNA Gene Probe Differentiates Member Species of the *Anopheles Gambiae* Complex." *The American Journal of Tropical Medicine and Hygiene* 37.1 (1987) 37–41.

Čolović, M. B. *et al*. "Acetylcholinesterase inhibitors: Pharmacology and toxicology." *Current neuropharmacology* 11.3 (2013): 315.

Coluzzi, M. *et al*. "Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex." *Transactions of the Royal Society of Tropical Medicine and Hygiene* 73.5 (1979): 483-497.

Coluzzi M. *et al*. "A polytene chromosome analysis of the *Anopheles gambiae* species complex." *Science* 298 (2002): 1415-1418.

Costantini, C. *et al*. "Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in *Anopheles gambiae*." *BMC Ecology* 9, 16 (2009).

Coyne, J. A. and H. A. Orr. "Speciation." Vol. 37. *Sunderland, MA: Sinauer Associates* (2004).

Cramer, D. "*Basic Statistics for Social Research"* (Routledge, London, 1997).

Crawford, J. E. and B. P. Lazzaro. "The demographic histories of the M and S molecular forms of *Anopheles gambiae s.s.*" *Molecular Biology and Evolution* 27.8 (2010): 1739-1744.

Cruickshank, T. E. and M. W. Hahn. "Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow." *Molecular Ecology* 23.13 (2014): 3133-3157.

Currie-Jordan, A. "Genetic Basis of Pyrethroid Insecticide Resistance in Natural Populations of *Anopheles arabiensis* from Eastern Uganda." Masters Thesis – *Liverpool School of Tropical Medicine* (2015)

Cutter, A. D. and B. A. Payseur. "Genomic signatures of selection at linked sites: unifying the disparity among species." *Nature Reviews Genetics* 14 (2013): 262-274.

Dabiré, K. R. *et al.* "Distribution of pyrethroid and DDT resistance and the L1014F kdr mutation in *Anopheles gambiae s.l.* from Burkina Faso (West Africa)." *Transactions of the Royal Society of Tropical Medicine and Hygiene* 103 (2009): 1113–1120.

Dabiré, K. R. *et al.* "Assortative mating in mixed swarms of the mosquito *Anopheles gambiae s.s.* M and S molecular forms, in Burkina Faso, West Africa." *Medical Veterinary Entomology* 27 (2013): 298-312.

Danecek, P. *et al.* "The variant call format and VCFtools." *Bioinformatics* 27 (2011): 2156–2158.

Darwin, C. "On the origins of species by means of natural selection." *London: Murray* (1859).

Davidson, G. "The five mating-types in the *Anopheles Gambiae* complex." *Rivista di Malariologia* 43 (1964): 167.

Davies, T. G. E. *et al.* "DDT, pyrethrins, pyrethroids and insect sodium channels." *IUBMB Life* 59 (2007a): 151–162.

Davies, T. G. E. *et al.* "A comparative study of voltage-gated sodium channels in the Insecta: implications for pyrethroid resistance in Anopheline and other Neopteran species." *Insect Molecular Biology* 16.3 (2007b): 361-375.

Davis, N. A., A. Pandey, and B. A. McKinney. "Real-world comparison of CPU and GPU implementations of SNPrank: a network analysis tool for GWAS." *Bioinformatics* 27.2 (2011): 284-285.

Denholm, I., G. J. Devine and M. S. Williamson. "Insecticide resistance on the move." *Science* 297 (2002): 2222–2223.

Diabaté, A. *et al.* "Evaluating the effect of postmating isolation between molecular forms of *Anopheles gambiae* (Diptera: Culicidae)." *Journal of Medical Entomology* 44.1 (2007): 60-64.

Diabaté, A. *et al.* "Spatial swarm segregation and reproductive isolation between the molecular forms of *Anopheles gambiae*." *Proceedings of the Royal Society B.* 276 (2009): 4215–4222.

Dieckmann, U. and M. Doebeli. "On the origin of species by sympatric speciation." *Nature* 400.6742 (1999): 354-357.

Diehl, S. R. and G. L. Bush. "The role of habitat preference in adaptation and speciation." *Speciation and its Consequences* (1989): 345-365.

Djogbénou, L. *et al.* "Characterization of insensitive acetylcholinesterase (ace-1R) in *Anopheles gambiae* (Diptera: Culicidae): resistance levels and dominance." *Journal of Medical Entomology* 44.5 (2007): 805-810.

Djogbénou, L. *et al.* "Identification and geographic distribution of the ACE-1R mutation in the malaria vector *Anopheles gambiae* in south-western Burkina Faso, West Africa." *The American Journal of Tropical Medicine and Hygiene* 78.2 (2008a): 298-302.

Djogbénou, L. *et al.* "Evidence of introgression of the *ace-1*^{R} mutation and of the *ace-1* duplication in West African *Anopheles gambiae ss" PLoS ONE* 3.5 (2008b): e2172.

Djogbénou, L., V. Noel and P. Agnew. "Costs of insensitive acetylcholinesterase insecticide resistance for the malaria vector *Anopheles gambiae* homozygous for the G119S mutation." *Malaria Journal* 9.1 (2010): 12.

Dobzhansky, *T. "Studies on hybrid sterility. I. Spermatogenesis in pure and hybrid Drosophila pseudoobscura."* Zeitschrift für Zellforschung und Mikroskopische Anatomie 21 (1934)**:** 169–*221*.

Dobzhansky T. "Genetics and the Origin of Species*." Columbia University Press, New York* (1937).

Edi, C. V. *et al.* "CYP6 P450 enzymes and *ACE-1* duplication produce extreme and multiple insecticide resistance in the malaria mosquito Anopheles gambiae." *PLoS Genetics* 10.3 (2014): e1004236.

Eeles, R. A. *et al.* "Multiple newly identified loci associated with prostate cancer susceptibility." *Nature Genetics* 40.3 (2008): 316-321.

Ellegren, H. *et al*. "The genomic landscape of species divergence in *Ficedula* flycatchers." *Nature* 491.7426 (2012): 756-760.

Essandoh, J., A. E. Yawson and D. Weetman. "Acetylcholinesterase (Ace-1) target site mutation 119S is strongly diagnostic of *Anopheles gambiae* carbamate and organophosphate resistance across southern Ghana." *Malaria Journal* 12, 404 (2013).

Etang, J. *et al.* "Polymorphism of intron-1 in the voltage-gated sodium channel gene of *Anopheles gambiae* s.s. populations from Cameroon with emphasis on insecticide knockdown resistance mutations." *Molecular Ecology* 18 (2009): 3076-3086.

Ewing, G. and J. Hermisson. "MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus." *Bioinformatics* 26.16 (2010): 2064-2065.

Fay, J. C. and C-I. Wu. "Sequence divergence, functional constraint, and selection in protein evolution." *Annual Review of Genomics and Human Genetics* 4.1 (2003): 213-235.

Favia, G. *et al*. "Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of *Anopheles gambiae* ss." *Insect Molecular Biology* 10.1 (2001): 19-23.

Feder, J. L. *et al*. "The effects of winter length on the genetics of apple and hawthorn races of *Rhagoletis pomonella* (Diptera: Tephritidae)." *Evolution* (1997): 1862-1876.

Feder, J. L. "The apple maggot fly, *Rhagoletis pomonella*." *Endless forms: species and speciation. Oxford Univ. Press, New York* (1998): 130-144.

Feder, J. L. *et al*. "Establishment of new mutations under divergence and genome hitchhiking." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367.1587 (2012): 461-474.

Feder, J. L., S. P. Egan and P. Nosil. "The genomics of speciation-with-gene-flow." *Trends in Genetics* 28.7 (2012): 342-350.

Ferrarini, M. *et al*. "An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome." *BMC Genomics* 14.1 (2013): 670.

Flaxman, A. D. *et al*. "Rapid scaling up of insecticide-treated bed net coverage in Africa and its relationship with development assistance for health: a systematic synthesis of supply, distribution, and household survey data." *PLoS Medicine* 7.8 (2010): 1011.

Flint, J. and E. Eskin. "Genome-wide association studies in mice." *Nature Reviews Genetics* 13.11 (2012): 807-817.

Fontaine, M. C. *et al*. "Extensive introgression in a malaria vector species complex revealed by phylogenomics." *Science* 347.6217 (2015): 1258524.

Foster, S. P., *et al*. "Analogous pleiotropic effects of insecticide resistance genotypes in peach–potato aphids and houseflies." *Heredity* 91.2 (2003): 98-106.

ffrench-Constant, R. H., P. J. Daborn and G. Le Goff. "The genetics and genomics of insecticide resistance." *Trends in Genetics* 20.3 (2004): 163-170.

ffrench-Constant, R. H. "Which came first: insecticides or resistance?" *Trends in Genetics* 23.1 (2007): 1-4.

Futschik, A. and C. Schlötterer. "The next generation of molecular markers from massively parallel sequencing of pooled DNA samples." *Genetics* 186.1 (2010): 207-218.

García, G. P. *et al*. "Recent rapid rise of a permethrin knock down resistance allele in *Aedes aegypti* in Mexico." *PLoS Neglected Tropical Diseases* 3.10 (2009): e531-e531.

Gavrilets, S. "Perspective: models of speciation: what have we learned in 40 years?" *Evolution* 57.10 (2003): 2197-2215.

Gentile, G. *et al*. "Attempts to molecularly distinguish cryptic taxa in *Anopheles gambiae ss*." *Insect Molecular Biology* 10.1 (2001): 25-32.

Gillies, M. T. and B. DeMeillon. "The Anophelinae of Africa South of the Sahara (Ethiopian Zoogeographical Region)" 2nd. Johannesburg , South African Institute of Medical Research (1968).

Giraldo-Calderón, G. I. *et al*. "VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases." *Nucleic Acids Research* (2014).

Gordicho, V. *et al*. "First report of an exophilic *Anopheles arabiensis* population in Bissau City, Guinea-Bissau: recent introduction or sampling bias?" *Malaria Journal* 13.1 (2014): 423.

Gray, E. M. *et al*. "Inversion 2La is associated with enhanced desiccation resistance in *Anopheles gambiae*." *Malaria Journal* 8 (2009): 215.

Griffin, J. T. *et al*. "Reducing Plasmodium falciparum malaria transmission in Africa: a model-based evaluation of intervention strategies." *PLoS Medicine* 7.8 (2010): 1028.

Gudmundsson, J. *et al*. "Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes." *Nature Genetics* 39.8 (2007): 977-983.

Haldane, J. BS. "Sex ratio and unisexual sterility in hybrid animals." *Journal of Genetics* 12.2 (1922): 101-109.

Hamblin, M. T. and C. F. Aquadro. "High nucleotide sequence variation in a region of low recombination in *Drosophila simulans* is consistent with the background selection model." *Molecular biology and evolution* 13.8 (1996): 1133-1140.

Hansen, T. F. "Why epistasis is important for selection and adaptation." *Evolution* 67 (2013): 3501-3511.

Hedrick, P. W. "Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation." *Molecular Ecology* 22 (2013): 4606-4618.

Hemingway, J. "The molecular basis of two contrasting metabolic mechanisms of insecticide resistance." *Insect Biochemistry and Molecular Biology* 30.11 (2000): 1009-1015.

Hemingway, J. *et al*. "The molecular basis of insecticide resistance in mosquitoes." *Insect Biochemistry and Molecular Biology* 34.7 (2004): 653-665.

Hennig, W. "Phylogenetic systematics" *University of Illinois Press* (1966).

Hey, J. "Recent advances in assessing gene flow between diverging populations and species." *Current Opinion in Genetics & Development* 16.6 (2006): 592-596.

Hohenlohe, P. A. *et al*. "Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367.1587 (2012): 395-408.

Holt, R. A. *et al*. "The genome sequence of the malaria mosquito *Anopheles gambiae*." *Science* 298.5591 (2002): 129-149.

Hunter, D. J. *et al*. "A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer." *Nature Genetics* 39.7 (2007): 870-874.

Iliadis, A., D. Anastassiou and X. Wang. "Fast and accurate haplotype frequency estimation for large haplotype vectors from pooled DNA data." *BMC Genetics* 13.1 (2012): 94.

Jacob, F. "Evolution and tinkering." *Science* 196 (1977): 1161-1166.

Joanest, D. N. and C. A. Gill. “Comparing measures of sample skewness and kurtosis.” *Statistician* 47 (1998): 183-189.

Jones, C. M. *et al*. “Footprints of positive selection associated with a mutation (N1575Y) in the voltage-gated sodium channel of *Anopheles gambiae.” Proceedings of the National Academy of Sciences* 109 (2012a): 6614-6619.

Jones, C. M. *et al*. "Additional selection for insecticide resistance in urban malaria vectors: DDT resistance in *Anopheles arabiensis* from Bobo-Dioulasso, Burkina Faso." (2012b): e45995.

Jones, C. M. *et al*. "The dynamics of pyrethroid resistance in *Anopheles arabiensis* from Zanzibar and an assessment of the underlying genetic basis." *Parasite and Vectors* 6.1 (2013): 343.

Joron, M. *et al*. "A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies." *PLoS Biology* 4.10 (2006): e303.

Kabula, B. *et al*. "Susceptibility status of malaria vectors to insecticides commonly used for malaria control in Tanzania." *Tropical Medicine & International Health* 17.6 (2012): 742-750.

Karasov, T., P. W. Messer and D. A. Petrov. "Evidence that adaptation in *Drosophila* is not limited by mutation at single sites." *PLoS Genetics* 6.6 (2010): e1000924.

Karlsen, B. O. *et al*. "Genomic divergence between the migratory and stationary ecotypes of Atlantic cod." *Molecular Ecology* 22.20 (2013): 5098-5111.

Kimura, M. “The neutral theory of molecular evolution.” *Cambridge University Press* (1984).

Kirkpatrick, M. and N. Barton. "Chromosome inversions, local adaptation and speciation." *Genetics* 173.1 (2006): 419-434.

Kitau, J. *et al*. "Species shifts in the *Anopheles gambiae* complex: do LLINs successfully control Anopheles arabiensis." *PLoS One* 7.3 (2012): e31481.

Kofler, R., R. V. Pandey and C. Schlötterer. "PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq)." *Bioinformatics* 27.24 (2011): 3435-3436.

Korte, A. and A. Farlow. "The advantages and limitations of trait analysis with GWAS: a review." *Plant Methods* 9.1 (2013): 29.

Kronforst, M. R., *et al*. “Hybridization reveals the evolving genomic architecture of speciation.” *Cell Reports* 5 (2013): 666-677 (2013).

Krzywinski, M. *et al*. "Circos: an information aesthetic for comparative genomics." *Genome Research* 19.9 (2009): 1639-1645.

Lawniczak, M. K. N. *et al*. "Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences." *Science* 330.6003 (2010): 512-514.

Lawson, D. *et al*. "VectorBase: a data resource for invertebrate vector genomics." Nucleic Acids Research 37 (2009).

Lee Y, *et al*. “Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, *Anopheles gambiae*.” *Proceedings of the National Academy of Sciences* 110 (2013a): 19854-19859.

Lee, Y. *et al*. "Chromosome inversions, genomic differentiation and speciation in the African malaria mosquito *Anopheles gambiae*." *PloS One* 8.3 (2013b): e57887.

Lehmann, T. and A. Diabaté. "The molecular forms of Anopheles gambiae: a phenotypic perspective." *Infection, Genetics and Evolution* 8.5 (2008): 737-746.

Li, R. *et al*. "ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun." *PLoS Computational Biology* 1.4 (2005): e43.

Li, H., J. Ruan and R. Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores." *Genome Research* 18.11 (2008): 1851-1858.

Li, H. *et al*. “The Sequence Alignment / Map Format and SAMtools.” *Bioinformatics* 25 (2009): 2078-2079.

Li, H. and R .Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform.” *Bioinformatics* 26 (2009): 589-595.

Librado, P. and J. Rozas. “DnaSP v5: A software for comprehensive analysis of DNA polymorphism data.” *Bioinformatics* 25 (2009): 1451-1452.

Liu, N. *et al*. "Behavioral change, physiological modification, and metabolic detoxification: mechanisms of insecticide resistance." *Acta Entomologica Sinica* 49.4 (2006): 671.

Long, Q. *et al*. "PoolHap: inferring haplotype frequencies from pooled samples by next generation sequencing." *PloS One* 6.1 (2011): e15292.

Loughney, K., R. Kreber, and B. Ganetzky. "Molecular analysis of the para locus, a sodium channel gene in *Drosophila*." *Cell* 58.6 (1989): 1143-1154.

Lowry, D. B. "Landscape evolutionary genomics." *Biology Letters* (2010).

Lyimo, I. N. and H. M. Ferguson. "Ecological and evolutionary determinants of host species choice in mosquito vectors." *Trends in Parasitology* 25.4 (2009): 189-196.

Lynd, A. *et al*. "Field, genetic, and modeling approaches show strong positive selection acting upon an insecticide resistance mutation in *Anopheles gambiae* ss." *Molecular Biology and Evolution* 27.5 (2010): 1117-1125.

McCart, C. and R. ffrench-Constant. Dissecting the insecticide-resistance-associated cytochrome P450 gene Cyp6g1. *Pest Management Science* 64.6 (2008): 639-645.

McCoy, R. C. *et al*. "Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements." *PLoS one* (2014): e106689.

McLaren, W. *et al*. "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor." *Bioinformatics* 26.16 (2010): 2069-2070.

Manolio, T. A. *et al*. "Finding the missing heritability of complex diseases." *Nature* 461.7265 (2009): 747-753.

Manske, H. M. and D. P. Kwiatkowski. “LookSeq: a browser-based viewer for deep sequencing data.” *Genome Res*earch 19 (2009): 2125–2132.

Marsden, C. D. *et al*. "Diversity, differentiation, and linkage disequilibrium: prospects for association mapping in the malaria vector *Anopheles arabiensis*." *G3: Genes| Genomes| Genetics* 4.1 (2014): 121-131.

Martinez-Torres, D. *et al*. "Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector *Anopheles gambiae ss*." *Insect Molecular Biology* 7.2 (1998): 179-184.

Matowo, J. *et al*. "Genetic basis of pyrethroid resistance in a population of *Anopheles arabiensis*, the primary malaria vector in Lower Moshi, north-eastern Tanzania." *Parasites and Vectors* 7.1 (2014): 274.

Mawejje, H. D. *et al*. "Insecticide resistance monitoring of field-collected *Anopheles gambiae sl* populations from Jinja, eastern Uganda, identifies high levels of pyrethroid resistance." *Medical and Veterinary Entomology* 27.3 (2013): 276-283.

Mayr, E. "Systematics and the origin of species, from the viewpoint of a zoologist." *Harvard University Press* (1942).

Mayr, E. "Animal Species and Evolution." *Cambridge, Massachusetts: Belknap Press of Harvard University* (1963).

Megy, K. *et al*. "VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics." *Nucleic Acids Research* 40 (2012): 729-734.

Messer, P. W. and D. A. Petrov. "Population genomics of rapid adaptation by soft selective sweeps." *Trends in Ecology & Evolution* 28.11 (2013): 659-669.

Meyrowitsch, D. W. *et al*. "Is the current decline in malaria burden in sub-Saharan Africa due to a decrease in vector population." *Malaria Journal* 10.188 (2011): 10-1186.

Michel, A. P. *et al*. "Widespread genomic divergence during sympatric speciation." *Proceedings of the National Academy of Sciences* 107.21 (2010): 9724-9729.

Mitchell, S. *et al*. "Identification and validation of a gene causing cross-resistance between insecticide classes in *Anopheles gambiae* from Ghana." *Proceedings of the National Academy of Sciences* 109 (2012): 6147-6152.

Muller H. J. "Isolating mechanisms, evolution, and temperature." *Biology Symposium* 6 (1942): 71–125.

Müller, P. *et al*. "Field-caught permethrin-resistant *Anopheles gambiae* overexpress CYP6P3, a P450 that metabolises pyrethroids." *PLoS Genetics* 4.11 (2008): e1000286.

N'Guessan, R. *et al*. "Reduced efficacy of insecticide-treated nets and indoor residual spraying for malaria control in pyrethroid resistance area, Benin." *Emerging Infectious Diseases* 13.2 (2007): 199.

Nadeau, N. J. *et al*. "Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1587 (2012): 343-353.

Neafsey, D. E. *et al*. "SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes." *Science* 330.6003 (2010): 514-517.

Neafsey, D. E. *et al.* "Highly evolvable malaria vectors: The genomes of 16 *Anopheles mosquitoes*." *Science* 347.6217 (2015): 1258522.

Nei M (1987). Molecular Evolutionary Genetics. *Columbia University Press: New York*.

Noor, M. A. F. and S. M. Bennett. "Islands of speciation or mirages in the desert?; Examining the role of restricted recombination in maintaining species." *Heredity* 103.6 (2009): 439-444.

Nosil, P. "Speciation with gene flow could be common." *Molecular Ecology* 17.9 (2008): 2103-2106.

Nosil, P., D. J. Funk and D. Ortiz-Barrientos. "Divergent selection and heterogeneous genomic divergence." *Molecular Ecology* 18 (2009): 375–402.

Nosil, P. and J. L. Feder. "Genomic divergence during speciation: causes and consequences." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1587 (2012): 332-342.

Nwakanma, D. C. *et al.* "Breakdown in the process of incipient speciation in *Anopheles gambiae*." *Genetics* 193 (2013): 1221–1231.

Oh, S. *et al.* "Biochemical properties of recombinant acetylcholinesterases with amino acid substitutions in the active site." *Applied entomology and zoology* 42.3 (2007): 367-373.

Ohashi, J., I. Naka, and N. Tsuchiya. "The impact of natural selection on an ABCC11 SNP determining earwax type." *Molecular Biology and Evolution* 28.1 (2011): 849-857.

Ohta, Tomoko. "Role of gene duplication in evolution." *Genome* 31.1 (1989): 304-310.

Oliveira, E. *et al.* "High levels of hybridization between molecular forms of *Anopheles gambiae* from Guinea Bissau." *Journal of Medical Entomology* 45.6 (2008): 1057-1063.

Olsen, H. G. *et al.* "Genome-wide association mapping in Norwegian Red cattle identifies quantitative trait loci for fertility and milk production on BTA12." *Animal Genetics* 42.5 (2011): 466-474.

Park, J-H. *et al.* "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries." *Nature Genetics* 42.7 (2010): 570-575.

Parry, M. L. *et al.* "Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change" *Cambridge University Press*, Cambridge, UK (2007): 7-22.

Pates, H. and C. Curtis. "Mosquito behaviour and vector control." *Annual Review of Entomology*. 50 (2005): 53-70.

Patterson, N., A. L. Price, and D. Reich. "Population structure and eigenanalysis." *PLoS Genetics* 2.12 (2006): e190.

Pardo-Diaz, C. *et al*. "Adaptive introgression across species boundaries in *Heliconius* butterflies." *PLoS Genetics* 8 e1002752 (2012).

Pinto, J., *et al*. "An island within an island: genetic differentiation of Anopheles gambiae in Sao Tome, West Africa, and its relevance to malaria vector control." *Heredity* 91.4 (2003): 407-414.

Pinto, J. *et al*. "Multiple origins of knockdown resistance mutations in the Afrotropical mosquito vector *Anopheles gambiae*." *PLoS One* 2.11 (2007): e1243.

Poelstra, J. W. *et al*. "The genomic landscape underlying phenotypic integrity in the face of gene flow in crows." *Science* 344.6190 (2014): 1410-1414.

Pombi, M. *et al*. "Variation in recombination rate across the X chromosome of *Anopheles gambiae." Americal Journal of Tropical Medicine and Hygiene* 75 (2006): 901–903.

Pombi, M. *et al*. "Chromosomal plasticity and evolutionary potential in the malaria vector *Anopheles gambiae sensu stricto*: insights from three decades of rare paracentric inversions." *BMC Evolutionary Biology* 8.1 (2008): 309.

Posada, D. and K. A. Crandall. "Intraspecific gene genealogies: trees grafting into networks." *Trends in Ecology & Evolution* 16.1 (2001): 37-45.

Powell, T. H. Q. *et al*. "Genetic divergence along the speciation continuum: the transition from host race to species in *Rhagoletis* (Diptera: Tephritidae)." *Evolution* 67.9 (2013): 2561-2576.

Presgraves, D. C. "Sex chromosomes and speciation in *Drosophila*." *Trends in Genetics* 24, 336 (2008).

Price, A. L. *et al*. "Principal components analysis corrects for stratification in genome- wide association studies." *Nature Genetics* 38.8 (2006): 904-909.

Pritchard, J. K., M. Stephens and P. Donnelly "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2000): 945–59.

Pritchard, J. K., J. K. Pickrell, and G. Coop. "The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation." *Current Biology* 20.4 (2010): 208-215.

Pritchard, V. L. and S. Edmands. "The genomic trajectory of hybrid swarms: outcomes of repeated crosses between populations of *Tigriopus californicus*." *Evolution* 67.3 (2013): 774-791.

Quinlan, A. R. "BEDTools: The Swiss-Army Tool for Genome Feature Analysis." *Current Protocols in Bioinformatics* (2014): 11-12.

R Development Core Team R. "R: A Language and Environment for Statistical Computing." *R Foundation for Statistical Computing* 1, 409 (2011).

R Development Core Team. *"*R: A Language and Environment for Statistical Computing*."* *R Foundation for Statistical Computing*, Vienna*. (*2014*) Available at: http://www.R-project.org.*

Ranson, H. *et al.* "Identification of a point mutation in the voltage-gated sodium channel gene of Kenyan *Anopheles gambiae* associated with resistance to DDT and pyrethroids." *Insect Molecular Biology* 9.5 (2000): 491-497.

Ranson, H. *et al.* "Evolution of supergene families associated with insecticide resistance." *Science* 298.5591 (2002): 179-181.

Ranson, H. *et al.* "Insecticide resistance in *Anopheles gambiae*: data from the first year of a multi-country study highlight the extent of the problem." *Malaria Journal* 8.1 (2009): 299.

Ranson, H. *et al.* "Pyrethroid resistance in African anopheline mosquitoes: what are the implications for malaria control?" *Trends in Parasitology* 27.2 (2011): 91-98.

Reddy, M. R. *et al.* "Outdoor host seeking behaviour of *Anopheles gambiae* mosquitoes following initiation of malaria vector control on Bioko Island, Equatorial Guinea." *Malaria Journal* 10.1 (2011): 184.

Reidenbach, K. R. *et al.* "Patterns of genomic differentiation between ecologically differentiated M and S forms of *Anopheles gambiae* in West and Central Africa." *Genome Biology and Evolution* 4 (2012): 1202–1212.

O'Reilly, A., et al. "Modelling insecticide-binding sites in the voltage-gated sodium channel." *Biochemistry Journal* 396 (2006): 255-263.Reimer, L. J. *et al*. "An unusual distribution of the kdr gene among populations of *Anopheles gambiae* on the island of Bioko, Equatorial Guinea." *Insect Molecular Biology* 14.6 (2005): 683-688.

Renaut, S. *et al*. "Genomic islands of divergence are not affected by geography of speciation in sunflowers." *Nature Communications* 4, 1827 (2013).

Rice, W. R. and E. E. Hostert. "Laboratory experiments on speciation: what have we learned in 40 years?" *Evolution* (1993): 1637-1653.

Rinkevich, F. D., D. Yuzhe Du, and K. Dong. "Diversity and convergence of sodium channel mutations involved in resistance to pyrethroids." *Pesticide Biochemistry and Physiology* 106.3 (2013): 93-100.

Rocca, K. A. *et al*. "2La chromosomal inversion enhances thermal tolerance of *Anopheles gambiae* larvae." *Malaria Journal* 8 (2009): 147.

Ruegg, K. *et al*. "A role for migration-linked genes and genomic islands in divergence of a songbird." *Molecular Ecology* 23.19 (2014): 4757-4769.

Sabeti, P. C. *et al*. "Detecting recent positive selection in the human genome from haplotype structure." *Nature* 419.6909 (2002): 832-837.

Sabeti, P. C. *et al*. "Positive natural selection in the human lineage." *Science* 312.5780 (2006): 1614-1620.

Santolamazza, F. *et al*. "Insertion polymorphisms of SINE200 retrotransposons within speciation islands of *Anopheles gambiae* molecular forms." *Malaria Journal* 7, 163 (2008).

Santolamazza, F. *et al*. "Remarkable diversity of intron-1 of the para voltage-gated sodium channel gene in an Anopheles gambiae/Anopheles coluzzii hybrid zone." *Malaria Journal* 14.1 (2015): 1-10.

Scheet, P. and M. Stephens. "A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase." *American Journal of Human Genetics* 78 (2006): 629–644.

Schilthuizen, M., M. C. W. G. Giesbers, and L. W. Beukeboom. "Haldane's rule in the 21st century." *Heredity* 107.2 (2011): 95-102.

Schlötterer, C. *et al.* "Sequencing pools of individuals - mining genome-wide polymorphism data without big funding." *Nature Reviews Genetics* (2014).

Scott, J. A., W. G. Brogdon and F. H. Collins. "Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction." *American Journal of Tropical Medicine and Hygiene* 49 (1993): 520–529.

Seehausen, O. *et al.* "Genomics and the origin of species." *Nature Reviews Genetics* 15.3 (2014): 176-192.

Servedio, M. R. and M. A. F. Noor. "The role of reinforcement in speciation: theory and data." *Annual Review of Ecology, Evolution, and Systematics* (2003): 339-364.

Sharakhova, M. V. "Genome mapping and characterization of the *Anopheles gambiae* heterochromatin." *BMC Genomics* 11, 459 (2010).

Silva, A. P. B. *et al.* "Mutations in the voltage-gated sodium channel gene of anophelines and their association with resistance to pyrethroids–a review." *Parasites and Vectors* 7 (2014): 450.

Simard, F. *et al.* "Ecological niche partitioning between *Anopheles gambiae* molecular forms in Cameroon: the ecological side of speciation." *BMC ecology* 9.1 (2009): 17.

Sinka, M. *et al.* "The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurence, data, distribution maps and bionomic précis." *Parasites and Vectors* 3.117 (2010): 1-34.

Slotman M., A. della Torre and J. R. Powell. "Female sterility in hybrids between *Anopheles gambiae* and *A. arabiensis*, and the causes of Haldane's rule." *Evolution* 59 (2005): 1016-1026.

Slotman, M. A. *et al.* "Reduced recombination rate and genetic differentiation between the M and S forms of *Anopheles gambiae ss*." *Genetics* 174.4 (2006): 2081-2093.

Smadja, C., J. Galindo and R. Butlin. "Hitching a lift on the road to speciation." *Molecular Ecology* 17 (2008): 4177–4180.

Smith, J. M. "Sympatric speciation." *American Naturalist* (1966): 637-650.

Smoot, M. *et al.* "Cytoscape 2.8: new features for data integration and network visualization." *Bioinformatics* 27(3) (2010): 431–432.

Song, Y. *et al*. “Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice.” *Current Biology* 21 (2011): 1296-1301.

Sonoda, S. *et al*. "Genomic organization of the para-sodium channel α-subunit genes from the pyrethroid-resistant and-susceptible strains of the diamondback moth." *Archives of Insect Biochemistry and Physiology* 69.1 (2008): 1-12.

Sonoda, S. *et al*. "Duplication of acetylcholinesterase gene in diamondback moth strains with different sensitivities to acephate." *Insect Biochemistry and Molecular Biology* 48 (2014): 83-90.

Stephens, M., N. J. Smith, and P. Donnelly. “A new statistical method for haplotype reconstruction from population data.” *American Journal of Human Genetics* 68 (2001): 978–989.

Stephens, M. and P. Scheet. “Accounting for decay of linkage disequilibrium in haplotype inference and missing data imputation.” *American Journal of Human Genetics* 76 (2005): 449–462.

Stump, A. D. *et al*. "Centromere-proximal differentiation and speciation in *Anopheles gambiae*." *Proceedings of the National Academy of Sciences* 102.44 (2005): 15930-15935.

Sudia, W. D. and R. W. Chamberlain. “Battery-Operated Light Trap. An Improved Model.” *Mosquito News* 22: (1962) 126–29.

Tajima, F. “Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.” *Genetics* 123 (1989): 585–595.

Takahata, N. and M. Nei. “Gene genealogy and variance of interpopulational nucleotide differences.” *Genetics* 110 (1985): 325-344.

Tamura, K. *et al*. "MEGA6: molecular evolutionary genetics analysis version 6.0." *Molecular Biology and Evolution* 30.12 (2013): 2725-2729.

Templeton, A. R., K. A. Crandall and C. F. Sing. "A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation." *Genetics* 132.2 (1992): 619-633.

Toé, K. H. *et al*. "Increased pyrethroid resistance in malaria vectors and decreased bed net effectiveness, Burkina Faso." *Emerging Infectious Diseases* 20.10 (2014): 1691.

della Torre, A. *et al.* "Selective introgression of paracentric inversions between two sibling species of the *Anopheles gambiae* complex." *Genetics* 146.1 (1997): 239-244.

della Torre, A. *et al.* "Molecular evidence of incipient speciation within *Anopheles gambiae s.s.* in West Africa." *Insect Molecular Biology* 10 (2001): 9-18.

della Torre, A., Z. Tu, and P. Petrarca. "On the distribution and genetic differentiation of *Anopheles gambiae* ss molecular forms." *Insect biochemistry and molecular biology* 35.7 (2005): 755-769.

Tripet, F. *et al.* "DNA analysis of transferred sperm reveals significant levels of gene flow between molecular forms of *Anopheles gambiae*." *Molecular Ecology* 10.7 (2001): 1725-1732.

Tripet, F. *et al.* "Frequency of multiple inseminations in field-collected *Anopheles gambiae* females revealed by DNA analysis of transferred sperm." *The American Journal of Tropical Medicine and Hygiene* 68.1 (2003): 1-5.

Turner, S. D. "qqman: an R package for visualizing GWAS results using QQ and Manhattan plots." *bioRxiv* (2014): 005165.

Turner, T. L., M. W. Hahn, and S. V. Nuzhdin. "Genomic islands of speciation in *Anopheles gambiae*." *PLoS Biology* 3.9 (2005): 1572.

Turner, T. L. and M. W. Hahn. "Locus- and population-specific selection and differentiation between incipient species of *Anopheles gambiae." Molecular Biology and Evolution* 24 (2007): 2132–2138.

Turner, T. L. and M. W. Hahn. "Genomic islands of speciation or genomic islands and speciation?" *Molecular Ecology* 19.5 (2010): 848-850.

Turner, T. L., P. M. Miller and V. A. Cochrane. "Combining genome-wide methods to investigate the genetic complexity of courtship song variation in *Drosophila melanogaster*." *Molecular Biology and Evolution* 30.9 (2013): 2113-2120.

Via, S. A. C. Bouck and S. Skillman. "Reproductive isolation between divergent races of pea aphids on two hosts. II. Selection against migrants and hybrids in the parental environments." *Evolution* 54.5 (2000): 1626-1637.

Via, S. and J. West. "The genetic mosaic suggests a new role for hitchhiking in ecological speciation." *Molecular Ecology* 17 (2008): 4334–4345.

Via, S. "Natural selection in action during speciation." *Proceedings of the National Academy of Sciences* 106.Supplement 1 (2009): 9939-9946.

Via, S. "Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1587 (2012): 451-460.

Vontas, J. *et al*. "Gene expression in insecticide resistant and susceptible Anopheles gambiae strains constitutively or after insecticide exposure." *Insect Molecular Biology* 14.5 (2005): 509-521.

Vontas, J. *et al*. "Transcriptional analysis of insecticide resistance in *Anopheles stephensi* using cross-species microarray hybridisation." *Insect Molecular Biology* 16 (2007): 315-324.

Wahlund, S. "The combination of populations and the appearance of correlation examined from the standpoint of the study of heredity." *Hereditas* 11 (1928): 65-106.

Wakeley J. The variance of pairwise nucleotide differences in two populations with migration. *Theoretical Population Biology* 49 (1996) 39-57.

Weetman, D. *et al*. "Association mapping of insecticide resistance in wild *Anopheles gambiae* populations: major variants identified in a low-linkage disequilbrium genome." *PloS One* 5, e13140 (2010).

Weetman, D. *et al*. Gene flow-dependent genomic divergence between *Anopheles gambiae* M and S forms. *Molecular Biology and Evolution* 29 (2012): 279–291.

Weetman, D. *et al*. "Contemporary gene flow between wild *An. gambiae ss and An. arabiensis*." *Parasites and Vectors* 7 (2014): 345.

Weetman, D. *et al*. "Contemporary evolution of resistance at the major insecticide target site gene Ace-1 by mutation and copy number variation in the malaria mosquito *Anopheles gambiae*." *Molecular Ecology* 24.11 (2015): 2656-2672.

Weill, M. *et al*. The kdr mutation occurs in the Mopti form of *Anopheles gambiae s.s.* through introgression. *Insect Molecular Biology* 9 (2000): 451–455.

Weill, M. *et al*. "The unique mutation in *ace-1* giving high insecticide resistance is easily detectable in mosquito vectors." *Insect Molecular Biology* 13.1 (2004): 1-7.

Weill, M. *et al*. "The unique mutation in ace-1 giving high insecticide resistance is easily detectable in mosquito vectors." *Insect Molecular Biology* 13.1 (2004): 1-7.

Weir, B.S. and C. C. Cockerham. “Estimating F-statistics for the analysis of population structure.” *Evolution* 38 (1984): 1358–1370.

White, B. J. *et al*. "The population genomics of trans-specific inversion polymorphisms in *Anopheles gambiae*." *Genetics* 183.1 (2009): 275-288.

White, B. J. *et al*. "Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*." *Molecular Ecology* 19.5 (2010): 925-939.

Wilding, C. S., R. K. Butlin, and J. Grahame. "Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers." *Journal of Evolutionary Biology* 14.4 (2001): 611-619.

Wilding, C. S. *et al*. "High, clustered, nucleotide diversity in the genome of *Anopheles gambiae* revealed through pooled-template sequencing: implications for high-throughput genotyping protocols." *BMC Genomics* 10.1 (2009): 320.

Williamson, M. S. *et al*. “Identification of mutations in the housefly para-type sodium channel gene associated with knockdown resistance (*kdr*) to pyrethroid insecticides. *Molecular Genetics and Genomics* 252 (1996): 51-60.

Willing, E. M., C. Dreyer and C. Van Oosterhout. “Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers.” *PloS One* 7, e42649 (2012).

Witzig, C.*et al*. "Genetic mapping identifies a major locus spanning P450 clusters associated with pyrethroid resistance in *kdr*-free *Anopheles arabiensis* from Chad." *Heredity* 110.4 (2013): 389-397.

Wondji, C. S. *et al*. “Two duplicated P450 genes are associated with pyrethroid resistance in *Anopheles funestus*, a major malaria vector.” *Genomic Research* 19 (2009): 452-459.

World Health Organization. “WHO Global Malaria Programme - World Malaria Report.” Geneva (2011).

World Health Organization. “Global Plan for Insecticide Resistance Management (GPIRM).” Geneva (2012).

World Health Organisation. “Test procedures for insecticide resistance monitoring in malaria vector mosquitoes.” Geneva (2013).

World Health Organization. "World Malaria Report 2014.” Geneva (2014).

World Urbanization Prospects: the 2014 Revision [http:// http://esa.un.org/unpd/wup/Highlights/WUP2014-Highlights.pdf] Accessed September 24, 2015.

Wu, C-I. "The genic view of the process of speciation." *Journal of Evolutionary Biology* 14.6 (2001): 851-865.

Wu, C-I. and C-T Ting. "Genes and speciation." *Nature Reviews Genetics* 5.2 (2004): 114-122.

Xu, Q. *et al*. "Sodium channel genes and their differential genotypes at the L-to-F kdr locus in the mosquito *Culex quinquefasciatus*." *Biochemical and Biophysical Research Communications* 407.4 (2011): 645-649.

Yawson, A. E. *et al*. "Species abundance and insecticide resistance of *Anopheles gambiae* in selected areas of Ghana and Burkina Faso." *Medical Veterinary Entomology* 18 (2004): 372–377.

Yawson, A. E. *et al*. "Ecological zones rather than molecular forms predict genetic differentiation in the malaria vector *Anopheles gambiae s.s.* in Ghana." *Genetics* 175 (2007): 751-761.

Yoon, S. *et al*. "Sensitive and accurate detection of copy number variants using read depth of coverage." *Genome Research* 19.9 (2009): 1586-1592.